STATISTICAL POLICY
WORKING PAPER 36

Seminar on the Funding Opportunity

in Survey and Statistical Research

Federal Committee on Statistical Methodology

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

June 2004

The Federal Committee on Statistical Methodology
(June 2004)

Members

Brian A. Harris-Kojetin, Chair, Office of Management and Budget

Wendy L. Alvey, Secretary, U.S. Census Bureau

Lynda Carlson, National Science Foundation

Cynthia Z.F. Clark, U.S. Census Bureau

Steven B. Cohen, Agency for Healthcare Research and Quality

Stephen H. Cohen, Bureau of Labor Statistics

Lawrence H. Cox, National Center for Health Statistics

Zahava D. Doering, Smithsonian Institution

Robert E. Fay, U.S. Census Bureau

Ronald Fecso, National Science Foundation

Dennis Fixler, Bureau of Economic Analysis

Gerald Gates, U.S. Census Bureau

Barry Graubard, National Cancer Institute

Brian Greenberg, Social Security Administration

William Iwig, National Agricultural Statistics Service

Arthur Kennickell, Federal Reserve Board

Nancy J. Kirkendall, Energy Information Administration

Susan Schechter, Office of Management and Budget

Rolf R. Schmitt, Federal Highway Administration

Marilyn Seastrom, National Center for Education Statistics

Monroe G. Sirken, National Center for Health Statistics

Nancy L. Spruill, Department of Defense

Clyde Tucker, Bureau of Labor Statistics

Alan R. Tupek, U.S. Census Bureau

G. David Williamson, Centers for Disease Control and Prevention

Expert Consultant

Robert Groves, Joint Program in Survey Methodology

Table of Contents

Session 1. Origins of the Funding Opportunity in Survey Research

Session 2. Bayesian Methodology for Disclosure Limitation and Statistical Analysis of Large Government Surveys

Session 3. Visual and Interactive Issues in the Design of Web Surveys

Session 4. Robust Small Area Estimation Based on a Survey Weighted MCMC Solution for the General Linear Mixed Model

Session 5. Small Area and Longitudinal Estimation Using Information from Multiple Surveys

Session 6. Benefits and Challenges of the Funding Opportunity

Program on the Funding Opportunity in Survey Research Seminar

Bureau of Labor Statistics

2 Massachusetts Avenue, NE Washington, DC 20212

June 9, 2003

9:00 a.m. –

Welcoming Remarks

Nancy Kirkendall, Energy Information Administration, DOE

9:05 a.m. –

Session 1. Origins of the Funding Opportunity in Survey Research

Chair: Nancy Kirkendall, Energy Information Administration, DOE

Speaker: Monroe Sirken, National Center for Health Statistics

9:30 a.m. –

Session 2. Bayesian Methodology for Disclosure Limitation and Statistical Analysis of Large Government Surveys

Chair: Kathleen Utgoff, Bureau of Labor Statistics

Speakers: Roderick J. Little and Trivellore Raghunathan, University of Michigan

Discussants: Ramesh Dandekar, Energy Information Administration, DOE
William Winkler, U.S. Census Bureau

10:45 a.m. –

Session 3. Visual and Interactive Issues in the Design of Web Surveys

Chair: Susan Grad, Social Security Administration

Speaker: Roger Tourangeau, University of Michigan

Discussants: John Bosley, Bureau of Labor Statistics

Cleo Redline, National Science Foundation

1:00 p.m. –

Session 4. Robust Small Area Estimation Based on a Survey Weighted MCMC Solution for the General Linear Mixed Model

Chair: Marilyn Seastrom, National Center for Educational Statistics

Speakers: Ralph E. Folsom and Avinash Singh, Research Triangle Institute

Discussants: William Davis, National Cancer Institute

Phil Kott, National Agricultural Research Service

2:00 p.m. –

Session 5. Small Area and Longitudinal Estimation Using Information from Multiple Surveys

Chair: Lynda Carlson, National Science Foundation

Speaker: Sharon Lohr, Arizona State University

Discussants: Charles Perry, National Agricultural Statistical Research Service

John Eltinge, Bureau of Labor Statistics

3:00 p.m. –

Session 6. Benefits and Challenges of the Funding Opportunity

Chair: Nancy Kirkendall, Energy Information Administration, DOE

Discussants: Fritz Scheuren, National Opinion Research Center

Brian Harris-Kojetin, Office of Management and Budget

4:00 p.m. - **Adjourn**

Session 1

Origins of the Funding Opportunity in Survey Research

Charting the Interdisciplinary History of the Funding Opportunity in Survey and Statistical Research

Monroe G. Sirken
National Center for Health Statistics

The Funding Opportunity in Survey and Statistical Research is an interdisciplinary grants program in basic survey and statistical research that is oriented to the needs of Federal statistical agencies. It was officially established in 1998 when the National Science Foundation (NSF), the Interagency Committee on Statistical Policy (ICSP)*, and the Federal Committee on Statistical Methodology (FCSM)** agreed to jointly fund and administer the program. However, the Funding Opportunity's heritage goes back much further than when it was created five years ago.

The Funding Opportunity is rooted in the CASM Movement. As a matter of fact, it would never have been established in 1998 or since, had it not been for the efforts underway in 1998 to obtain the funding needed to sustain the research agenda of the CASM Movement. The CASM Movement, a long-term effort to foster interdisciplinary research on the cognitive aspects of survey methodology, emerged in the early 1980's as a direct consequence of the change from the behavioral to the cognitive paradigm that occurred in psychology the early 1970's. Thus, charting the interdisciplinary history of the Funding Opportunity is equivalent to recounting the history of a sustained effort to foster interdisciplinary survey research in the United States that began more than 30 years ago, and is alive and active today due to the recent extension of the 1998 NSF/FCSM-ICSP agreement to fund and administer the Funding Opportunity in Survey and Statistical Research beyond the year 2002.

Historical Overview

The history of the Funding Opportunity in Survey and Statistical Research can be divided into the three periods shown below:

Period (1) The Prologue - The decade between the emergence of the cognitive paradigm in the early 1970's and convening the CASM I Seminar 1983;
Period (2) The CASM Movement - The 14-year period between the CASM I Seminar in 1983 and CASM II Seminar in 1997;
Period (3) The Funding Opportunity Program - The 5-year period since the Funding Opportunity in Survey and Statistical Research was established in 1998.

* The ICSP is a committee of the directors of 13 largest Federal statistical agencies and is chaired by the Chief Statistician of the Office of Management and Budget.

* * The FCSM is an interagency committee dedicated to improving the quality of Federal statistics and includes invited Federal agency staff with relevant experience and expertise. The reader is referred to Aborn (1999) for more information about period (1), to Tanur (1999) and Jabine (1999) for more information about Period (2), and to Sirken (2001) and Kirkendall (2001) for more information about Period (3).

The flowchart on the following page lists eight milestones in the history of the Funding Opportunity by the period of occurrence. Milestones occurring during periods 1, 2, and 3 respectively are discussed below. In view of the vital importance of interdisciplinary research to the advancement of official statistics, concluding section F proposes that studies should be undertaken to explore improved ways of meeting the challenges of fostering interdisciplinary research in our decentralized Federal Statistical System.

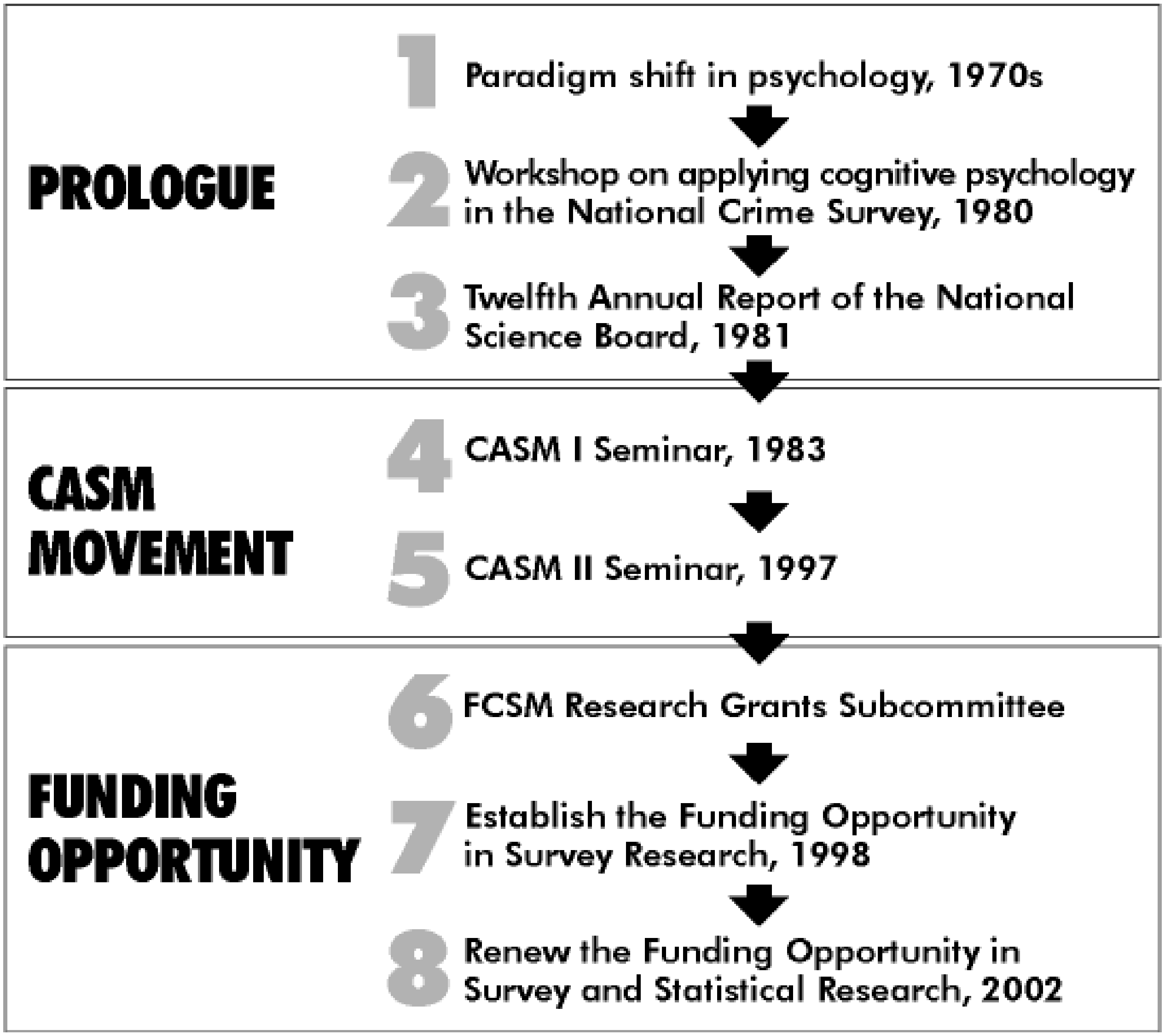
Milestones Preceding the CASM Movement

Milestone 1. Emergence of the cognitive paradigm. The shift in paradigms, from the behavioral to the cognitive, implied that the two-stage stimulus response process postulated by the behavioral paradigm is intervened by a cognitive stage in which subjects perform a range of mental tasks. Compared to the behavioral paradigm, the cognitive paradigm focuses on how the mind works rather than who the subjects are and what the conditions are when the subjects perform their tasks. The cognitive paradigm rapidly diffused from psychology and influenced the research orientations of other disciplines, notably education and computer science. When the cognitive paradigm reached the survey research community toward the end of the 1970's, it provided survey researchers and cognitive psychologist with opportunities to simultaneously address chronic measurement problems in survey response and non response, and to test in the real world of survey research, the cognitive theories that had been largely developed and tested in laboratory settings.

Consider, for example, the difference between the behavioral and cognitive theories of truthfully answering sensitive questions in surveys. Based on behavioral theory, the likelihood of truthful response depends on the survey takers' assessments of the sensitivity of the survey questions and the extent of privacy and anonymity provided by the data collection modes. Based on cognitive theory, the independent variables are the respondents' perceptions of the risks and losses of truthful disclosure. From the cognitive theory perspective, the likelihood of truthful survey response is an example of "human decision making under conditions of uncertainty' - a scientific field of enquiry that has long interested cognitive psychologists and for which Daniel Kahneman, a cognitive scientist, recently shared a Nobel prize in economics. Conducting interdisciplinary research on the cognitive aspects of truthful response to sensitive survey questions potentially benefits survey researchers and cognitive psychologists. It provides survey researchers with innovative theories of survey response and non response that are in the mainstream of mathematical statistics and modern science, and it provides cognitive psychologists with opportunities to test cognitive theories of human decision making under conditions of uncertainty in the real world venue of survey taking.

Figure 1

Milestones in the History of the Funding Opportunity in Survey and Statistical Research


PROLOGUE
1 Paradigm shift in psychology, 1970s
2 Workshop on applying cognitive psychology in the National Crime Survey, 1980
3 Twelfth Annual Report of the National Science Board, 1981
CASM MOVEMENT
4 CASM I Seminar, 1983
5 CASM II Seminar, 1997
FUNDING OPPORTUNITY
6 FCSM Research Grants Subcommittee
7 Establish the Funding Opportunity in Survey Research, 1998
8 Renew the Funding Opportunity in Survey and Statistical Research, 2002

Milestone 2. The workshop on applying cognitive psychology to recall problems of the National Crime Survey. The initial meetings of cognitive psychologists and survey researchers were independently convened in the UK in 1978 (Moss and Goldstein, 1979) and in the USA in 1980 (Biderman, Cantor, Lynch, and Martin, 1986) to discuss the cognitive aspects of retrospective reporting in single time population surveys. The U.S. meeting, a 2-day workshop convened by the Bureau of Social Science Research (BSSR) with support of the Bureau of Justice Statistics and the Bureau of the Census, brought together a small number of cognitive psychologists and survey researchers to discuss cognitive methods to improve recall of victimization in the National Crime Survey. (The Crime Survey asks respondents to retrospectively report incidents in which they were crime victims.) Though the workshop was not organized with the intent of fostering an interdisciplinary survey research movement, it made lasting impressions on those in attendance and several of them subsequently became key players in the CASM Movement.

Milestone 3. The Twelfth Annual Report of the National Science Board. In its annual report to the President in 1981, the National Science Board (NSB), the governing body of the NSF, appraised six areas in which basic research has significantly impacted society, and luckily as it turned out "survey research and opinion polls" was one of these areas. After describing the growth of survey research and polling in our modern society, the NSB report stresses the need for more research and refinement in measuring the behavioral and social dimensions of survey taking so that surveys can continue to benefit society in the future. The twelfth NSB report was very influential in setting NSF priorities during the 1980's Quoting Murray Aborn (1999), then head of NSF's Measurement Methods and Data Improvement (MMDI) Program, "... it is no exaggeration to say that the [NSB] report was instrumental in obtaining the budgetary increments that made it possible [for the MMDI program] to support the CASM I [Seminar], and subsequent research projects, and the cognitive research laboratory at the National Center for Health Statistics." Dr. Aborn was a consultant to the NSB in preparing the 12th annual report, and had been a participant at the BSSR Workshop.

Milestones of the CASM Movement

Milestone 4. The CASM I Seminar. The Advanced Research Seminar on the Cognitive Aspects of Survey Methodology, more familiarly known as the CASM I Seminar, convened on June 15, 1983. It was a landmark event that initiated the CASM Movement to foster interdisciplinary research on the cognitive aspects of survey methodology (Jabine et al, 1984). The CASM Seminar and its follow-up meeting in January 1984 were organized and convened by the Committee on National Statistics and funded by the MMDI program. Twenty-two invited cognitive psychologists and survey researchers from academia and government participated in the seminar and its follow-up meeting. The seminar sought to foster dialogues among the participants, and to develop ideas for collaborative research project proposals. Its success in realizing both objectives can be attributed to careful planning by CNSTAT staff, and to the announcement by the MMDI program before the seminar that it would be interested in funding the most promising seminar research project proposals. Several ideas for interdisciplinary research projects evolved at the CASM I Seminar, were discussed at the follow-up January meeting, and later were submitted by seminar participants to and were funded by the MMDI program

Milestone 5. The CASM II Seminar. The CASM II Seminar convened in June 1997 (Sirken et al, 1999b). The seminar was jointly funded by the NSF and National Center for Health Statistics (NCHS), organized by a Planning Committee of survey researchers from government and universities, and administered by NCHS staff. Though it had been prudent at the CASM I Seminar to narrowly focus CASM research on the cognitive aspects of questionnaire design, and to limit collaborations largely to cognitive psychologists, much had been accomplished since then and the CASM II Seminar sought to expand the scope of CASM research to address issues at all stages of the survey measurement process, and to expand collaborations to many scientific disciplines. The 6-day seminar had 58 invited participants, and 16 commissioned papers were presented and discussed (Sirken, et al., 1999a).

The CASM I Seminar served as the model for organizing of the CASM II Seminar. The CASM I and II Seminars were virtually equivalent in all major respects, except one. Unlike the CASM I Seminar, the CASM II Seminar lacked institutional funding to support the research projects generated at the CASM II Seminar and to sustain the CASM Movement thereafter. The sheltered CASM funding and administrative support previously provided by the MMDI program had expired with Murray Aborn's retirement in about 1990, and ongoing efforts to obtain commitments from the NSF and the Social Science Research Council to support CASM II research projects and to sustain the CASM Movement were in limbo when the CASM II Seminar convened. A potential break-through occurred towards the end of the CASM II Seminar when NSF's Methodology, Measurement, and Statistics (MMS) Program offered to administer a grants program in basic survey research and provide $300,000 in annual sheltered funding during a 3-year period, but the offer was contingent on matching funds being provided by a consortium of Federal Agencies.

Milestones of the Funding Opportunity

Milestone 6. Establishment of the FCSM Research Subcommittee. Well before the MMS program offer at the CASM II Seminar to cosponsor a survey research grants program, efforts had been underway to recruit and organize a consortium of Federal statistical agencies to support an interdisciplinary grants program in CASM research. In early 1997, when the matter was first brought to the attention of the Federal Committee on Statistical Methodology (FCSM), an FCSM Research Grants Subcommittee was appointed to draft a CASM II Research Consortium Proposal requesting the ICSP for concept approval and funding support for the consortium. The FCSM Subcommittee estimated that, at a minimum, about $600,000 or an average of almost $50,000 per ICSP agency (if all 13 ICSP agencies participated) would be required annually to maintain the CASM research grants program. Informal discussions with some ICSP agency heads made it clear that the proposed annual contribution of almost $50,000 per agency was unrealistic. However, the MMS offer to co-fund a survey research grants program would lower the average annual ICSP agency contribution from $50,000 to $25,000, and that reduction appeared to make the consortium proposal feasible.

Milestone 7. Establishment of the Funding Opportunity in Survey Research. In June 1998, the FCSM Research Subcommittee submitted a research grants program proposal to the ICSP with the following provisions: (1) the consortium of ICSP agencies match the MMS offer and contribute $300,000 annually for a period of three years to support meritorious research proposals of potential benefit to Federal statistical agencies; (2) ICSP agencies and NSF programs have opportunities to add-on funds for research proposals of particular interest to their respective programs; (3) project proposals responding to the MMS announcements undergo a two-tier project review and selection process, first by a NSF panel for scientific merit and then by a government panel for potential utility to Federal statistical agencies with final selections made by the MMS program in close collaboration with the FCSM Research Subcommittee; and (4) seminars, such as this one, that offer opportunities for direct discourse between the principal investigators of funded projects and statistical agency staffs are convened periodically in the Washington DC vicinity.

Twelve of the 13 ICSP agencies pledged to match the NSF contributions, by each contributing $25,000 annually for 3 years contingent on a successful demonstration during the first funding year. In September 1998, the ICSP, MMS and FCSM reached final agreement to administer and fund the Funding Opportunity during the 1999 demonstration year. In July 1999, the MMS/FCSM-ICSP agreement was extended for 2 additional funding years 2001 and 2002, and the name of the program was changed to The Funding Opportunity in Survey and Statistical Research.

Milestone 8. Renewal of the Funding Opportunity. Prior to the expiration in 2002 of the original NSF/FCSM-ICSP agreement, the MMS program indicated that if the ICSP agencies would continue to contribute $300,000 annually, MMS would be willing to continue to administer the Funding Opportunity, but as an integral part of the MMS grants program rather than a separate program. Also instead of the MMS program pledging $300,000 annually in sheltered funding for Funding Opportunity projects as it had in the past, the Funding Opportunity project proposals would compete on an equal basis with other project proposals submitted for MMS funding. Thus, in effect the size of the MMS contributions to the Funding Opportunity in the future would vary from year to year, and could be more or less than the $300,000 per annum contributed by the consortium of ICSP agencies.

During the latter part of 2002, the FCSM Research Committee incorporated the MMS renewal offer into a renewal proposal that was submitted to the ICSP with the recommendation that the ICSP agencies renew their pledges for 3 more years. The ICSP agencies agreed to extend their pledges and contribute a total of $300,000 annually to the Funding Opportunity for three more years, but instead of each ICSP agency contributing $25,000 annually, the sizes of the agency's annual contributions will vary somewhat depending on relative size of the agency's appropriated budget. The renewed NSF/FCSM-ICSP agreement became effective at the beginning of year 2003, using essentially the same NSF/FCSM administrative arrangements that had evolved and worked so well during the period of the first agreement.

Concluding Remarks

By its very nature survey research is an interdisciplinary discipline, and its advancement depends on knowledge and technology transfers that come about as a result of interdisciplinary survey research (Sirken and Herrmann, 1996). There is growing appreciation of the need to foster interdisciplinary survey research, but fostering interdisciplinary survey research is not an easy thing to do. Initiating interdisciplinary survey research requires bridging the communication and cultural gaps between survey researchers and researchers in other disciplines, and sustaining interdisciplinary research requires obtaining institutional commitments to provide the administrative structures and funding support (Olkin and Sacks, 1988). Fostering interdisciplinary survey research oriented to the needs of Federal statistical agencies is a particularly hard thing to do in our decentralized statistical system comprising 68 independent statistical agencies in which short term research linked to each agency's particular missions is the rule. Despite these difficulties or perhaps due to these difficulties, several independent efforts are currently underway to foster interdisciplinary research oriented to the needs of Federal statistical agencies. Other ongoing fostering efforts, in addition to the Funding Opportunity in Survey and Statistical Research, are the ASA/NSF Fellowship program and NSF's Digital Government Program.

In view of the vital importance of interdisciplinary research in the advancement of official statistics, it seems to me that initiating a research *project on the process of fostering interdisciplinary in official statistics* would be well worth the effort. As a step in that direction, I propose that a seminar be convened to review and compare the objectives and fostering strategies of the Funding Opportunity, the ASA/NSF Fellowship program, and NSF's Digital Government Program, and to discuss the policy implications of fostering interdisciplinary research efforts in the Federal statistical system.

Acknowledgements

The opinions expressed herein are those of the author and do not necessarily represent the official views or positions of the National Center for Health Statistics.

References

Aborn, Murray, (1999). In M. Sirken et al, Cognition and Survey Research (pp. 21-38). New York: Wiley.

Biderman, A., Cantor, D., Lynch, J., and Martin, E. (1986). Final report of the National Crime Survey Redesign Program. Washington, DC: Bureau of Social Science Research.

Jabine et.al. (1984). Cognitive Aspects of Survey Methdology: Building a Bridge Between Disciplines. Washington, DC National Academy Press.

Jabine, Thomas B., (1999). In M. Sirken et al (Eds), A New Agenda for Interdisciplinary Survey Research Methods (pp. 1-7). Hyattsville, MD: National Center for Health Statistics.

Kirkendall, Nancy (2001) In Statistical Policy Working Paper 33, Seminar on the Funding Opportunity In Survey Research. Washington, DC: Federal Committee on Statistical Methodology, OMB.

Moss, Louis and Goldstein, Harvey (1985) (Eds.). The Recall Method in Social Surveys. Portsmuth, NH: Heinemann.

National Science Foundation (1983). Only One Science: Twelfth Annual Report of the National Science Board: Washington, DC: Government Printing Office.

Olkin, Ingram, and Sachs, Jerome (1988). Cross-Disciplinary Research in the Statistical Sciences. Washington, DC: National Science Foundation.

Sirken, Monroe G. and Herrmann, Douglas (1996). Relationships Between Cognitive Psychology and Survey Research, Proceedings of the Social Statistics Section, American Statistical Association.

Sirken, Monroe G., Herman, Douglass, J., Schechter, Susan, Schwartz, Norbert, Tanur, Judith M., Tourangeau, Roger (1999a) (Eds.), Cognition and survey methods research. New York: Wiley.

Sirken, Monroe G., Jabine Thomas, Willis, Gordon, Martin, Elizabeth, Tucker, Clyde (1999b)(EDS.), A New Agenda for Interdisciplinary Survey Research Methods. Hyattsville, MD: Hyattsville: National Center for Health Statistics.

Sirken, Monroe G. (2001). In Statistical Policy Working Paper 33. Seminar on the Funding Opportunity in Survey Research (pp, 111-114) Washington, D.C., Federal Committee on Statistical Methodology, OMB.

Tanur, Judith M. (1999). In M. Sirken et al (Eds.), Cognition and Survey Research (pp. 13-19). New York: Wiley.

Session 2

Bayesian Methodology for Disclosure Limitation

and Statistical Analysis of Large Government Surveys

Bayesian Methodology for Disclosure Limitation and Statistical Analysis of Large Government Surveys

Discussant: Ramesh Dandekar, Energy Information Administration, U. S. DOE

Researchers: Roderick J. Little and Trivellore Raghunathan, University of Michigan

Background

Synthetic micro data has been used extensively to study the behavior of complex computer models for a long time. In recent years, there has been an increased realization that synthetic micro data could also be used for a dissemination of statistical information in place of real data containing sensitive records collected by federal agencies. Because of relatively low disclosure potential and the ability to recreate most of the statistical properties of the original data, synthetic micro data offers some advantage over other methods of micro data protection. It has also been known for a while that synthetic data offers an economical choice to the on-site data research centers operated by federal statistical agencies in dissemination of public use information. Ideally, potential researchers could use synthetic data from their own work site for initial hypothesis testing/model development, without concern for data confidentiality. The researchers will need to use the data center facility only to run their final refined model/setup on the original data. Such a strategy has the potential to reduce the on-site operating cost for data centers.

The characteristics of micro data disseminated by federal statistical agencies vary considerably. As a result, it is unlikely that one synthetic micro data generation method will work well on all different micro data types. This necessitates that statistical agencies conduct a broad-based research on multiple fronts to generate synthetic micro data. The two separate papers in this session offer unique application areas.

The paper by Raghunathan, Reiter and Rubin, “Multiple Imputation for Statistical Disclosure Limitation”, demonstrates the procedure to generate synthetic micro data by using multiple imputation framework proposed by Rubin in 1993. The proposed procedure uses a parametric and non parametric approach to generate synthetic data. The inference based on this technique requires that some adjustments be made to point and variance estimates prior to their use. The paper demonstrates that the inferences derived from the synthetic data are similar to those derived using actual data.

The Paper by Little and Liu, “Selective Multiple Imputation of Keys for Statistical Disclosure Control in Micro Data”, on the other hand, generates synthetic micro data by selective multiple imputation of categorical key variables and continuous non-key variables. The method offers a potential balance between data quality and statistical disclosure control by mixing select non-sensitive cases with sensitive cases.

Specific Comments

Both methods for synthetic micro data generation offer viable options by using a Bayesian framework. There are many potential applications for these two methods. However, the application potential for these two methods could be increased considerably by extending the scope of current research work to do the following:

1) Develop alternate methods/procedures to reduce current dependence on the model based imputation procedure. Developing the most appropriate global model to capture multi-variate statistical characteristics of any given data is always a time consuming process. It is also possible that the synthetic data end user might want to use the data to develop his/her own statistical model to represent original data. In such a situation, it might not be a good strategy to generate model-based synthetic data.

2) Derive new methods/procedures that will keep an optimum balance between the synthetic micro data quality and related tabular data quality along with adequate disclosure protection for both. It is a common practice to perform a preliminary statistical analysis of raw micro data by exploring associated tabular structure of the micro data. Conclusions derived from the tabular data analysis are commonly used in analytical studies and policy papers. Such a practice necessitates adequate precautions to retain statistical characteristics associated with original tabular structure to the extent possible.

3) Look at the feasibility of using the Latin Hypercube Sampling (LHS) method in combination with a restricted pairing algorithm by Iman and Conover to induce a desired rank correlation matrix on synthetic micro data within a framework supported by a Bayesian method. The LHS method is model independent and has been used successfully to generate synthetic micro data since late seventies. By using the empirical cumulative distribution function of the real data, the LHS method provides non-parametric approach to generate synthetic micro data. For many applications the LHS-based synthetic data generation method could offer the most practical approach that balances data quality and minimal resources required to generate synthetic micro data.

4) Look at the feasibility of performing backward calibration of micro data based on the outcome from the Controlled Tabular Adjustments (CTA) to protect related tabular data (Dandekar/Cox 2002, Dandekar 2003). Such a strategy allows one to one correspondence between synthetic micro data and synthetic tabular data.

References

Dandekar, Ramesh A. (1993), “Performance Improvement of Restricted Pairing Algorithm for Latin Hypercube Sampling”, ASA Summer conference (unpublished).

Dandekar Ramesh A. and Cox Lawrence H. (2002), Synthetic Tabular Data: An Alternative to Complementary Cell Suppression, manuscript available from ramesh.dandekar@eia.doe.gov.

Dandekar R. A., Cohen M., and Kirkendall, N. (2002a), Sensitive Micro Data Protection using Latin Hypercube Sampling Technique. In J. Domingo-Ferrer, ed., Inference Control in Statistical Databases, 117-125., Berlin:Springer-Verlag

Dandekar, R.A (2003), Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems, working paper 40, UNECE Work session on statistical data confidentiality (Luxembourg, 7-9 April 2003)

Iman R.L. and Conover W. J. (1982), "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables", Commun. Stat., B11(3): pp. 311-334.

McKay M.D., Conover W. J. and Beckman, R. J. (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code", Technometrics 21(2): pp. 239-245.

Discussion of

Multiple Imputation for Statistical Disclosure Limitation
by T. E. Raghunathan, J. P. Reiter, and D. B. Rubin

Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata
by R. J. A. Little and F. Liu

William E Winkler
U.S. Census Bureau

1. Introduction

Statistical agencies that provide public-use microdata must contend with the conflicting goals of producing data that satisfy one or more analytic needs of a group of users and preserving the confidentiality of data records associated with entities such as individuals or companies. It is the view of this discussant (e.g., Winkler 1997) that analytic needs should be met by building models of the public microdata. The models should be described in terms of user-specified requirements for analyses. The documentation should describe the limitations the microdata for the specified analytic purposes and other purposes to which the microdata might be put. If the analytic needs of the microdata have been justified, then the confidentiality of the microdata should have be described.

The outline of this discussion is as follows. In second section, I provide background on a number of existing methods and their analytic limitations. In the third section, I discuss the general framework of Raghunathan et al. (2003) for providing synthetic microdata under models that meet analytic needs and the framework of Little and Liu (2003) for providing partially synthetic data that also meets analytic needs and does not require the amount of modeling as the more general framework. The final section consists of concluding remarks.

2. Background

A variety of methods have been developed and used for masking a data file. The methods have the intent of altering the data in a manner that allows some analyses to be done that correspond to what could be done on the original, confidential microdata and of making re-identification more difficult. After masking, the resultant microdata are disseminated to users who presumably wish to perform analyses that could not be performed by using published tables alone.

These masking methods include swapping (Dalenius and Reiss 1982), rank swapping (Moore 1996), micro-aggregation (e.g., Domingo et al. 2002), k-similarity (Samarati and Sweeney 1998) that includes global recoding and local suppression, variants of additive noise (Kim 1986, 1990, Fuller 1993), and synthetic microdata (Rubin 1993, Fienberg 1997). All of the original and succeeding authors who have considered swapping, rank swapping, and micro-aggregation have been able to point out serious difficulties with providing for even basic analytic needs. If the swapping, rank swapping, and micro-aggregation are over relatively small and homogenous groups, then simple analytic needs may not be seriously compromised but re-identification can be straightforward (Winkler 2002). The method of Winkler (2002) for micro-aggregation it can be easily extended to swapping and rank swapping. Although k-similarity is guaranteed to provide confidentiality because at least k records with have the same identifying information, it has, so far, only been rigorously shown to provide analytic needs in very simple situations (Iyengar 2002). Sampling, as a simple alternative, neither assures that simple analytic needs are met nor assures that all records cannot be re-identified. Typically, sampling is not designed to satisfy a number of analytic constraints (particularly on a set of subdomains). With typical sampling designs, records in the sample can be population uniques and relatively straightforward to re-identify.

The only two methods that place primary emphasis on analytic properties of the masked microdata are the additive noise ideas of Kim (1986, 1990) and synthetic data methods (Rubin 1993, Fienberg 1997). A valid criticism of additive noise has been that it is only generally suitable for public-use microdata that is used in regression-type analyses. Another criticism has been that special software is needed for analyzing additive-noise microdata. High quality software (Yancey et al. 2002) is now available for correct analysis. The software even supports analyses on arbitrary subdomains according to the original ideas introduced by Kim (1990). At present, producing synthetic data according to models that consider user-specified analytic needs are the most promising approach. Criticisms of the approach deal with the inability of groups, particularly in statistical agencies, to develop models of their data and create software. A simplistic method for automatically creating models of the data using Bayesian networks was introduced by Thibaudeau and Winkler (2002). The standard methods for creating models for multiple imputation should still produce much higher quality analytic properties.

3. The Papers

This section summarizes and comments on the papers of Raghunathan, Reiter and Rubin (2003) and Little and Liu (2003).

3.1. Raghunathan, Reiter, and Rubin

The paper of Raghunathan et al. (2003) provides an important theoretical foundation for producing synthetic microdata satisfying analytic constraints. Three examples give insight and provide further practical advice. Other examples have been given by Reiter (2002, 2003). Further, software (Raghunathan et al. 1998) can facilitate producing microdata in a manner that is consistent with ideas introduced by Kennickell (1997) and Abowd and Woodcock (2002).

Fienberg (1997) raised the following issue. If sufficient analytic constraints are placed on the synthetic microdata, then some of the synthetic microdata records may be very close to actual population records. This has the possibility of allowing re-identification. In the Raghunathan et al. (2003) framework, arbitrary statistics q_M and T_M representing multiple imputation means and variances are considered. If a sufficiently large number of copies of the population P_i, $i \leq M$, are released and the models are sufficiently detailed to allow reasonable analyses on a moderate number of statistics q, when will it be possible that a moderate number of the original, confidential microdata records may be approximated with reasonable accuracy? Raghunathan et al. note that the approximate Bayesian bootstrap, while not as sensitive to model assumptions, can potentially lead to more re-identification. The parametric modeling, on the other hand, is more subject to model specification error as has been noted by Reiter (2002) in addition to Raghunathan et al..

3.2. Little and Liu

The paper of Little and Liu (2003) provides a practical framework for producing partially synthetic data that should be more straightforward to implement than purely synthetic data. Their paper also provides a useful and practical guide about how to do re-identification in straightforward situations.

I summarize their method. They assume that the original data consist of both continuous and discrete variables. They assume that outside individuals have a database that contains the discrete variables. Their method "masks" the discrete variables in a manner that does not change the continuous variables. They mask by choosing neighborhoods of variables using continuous variables only. Discrete data among "at risk" records or merely in a sample of records within the neighborhoods can be swapped. There is no requirement that the neighborhoods are disjoint. Their initial empirical results are promising. They demonstrate that the information loss due to the masking procedure is modest but still non-trivial. Using discrete data only, they provide re-identification risk metrics that are conservative and realistic.

If both continuous and discrete data are used for re-identification, is it possible to re-identify? Little and Liu might compare their information-loss/re-identification-risk framework to the R-U confidentiality map framework introduced by Duncan et al. 2001 (see also Trottini and Fienberg, 2002).

4. Concluding Remarks

The concluding remarks are two recommendations. The first recommendation is that all releases of public-use microdata should discuss and justify the analytic usefulness of the data. This should include what analyses on the original, confidential microdata can be reproduced on the masked, public-use microdata. The second recommendation is that the microdata confidentiality community should continue serious investigation of synthetic microdata, particularly with the information-loss/disclosure-risk framework given by both sets of authors. An alternative method for producing synthetic microdata using Latin Hypecubes is given by Dandekar et al. (2002).

References

Abowd, J. M. and Woodcock, S. D. (2002), "Disclosure Limitation in Longitudinal Linked Data," in (P. Doyle et al., eds.) *Confidentiality, Disclosure, and Data Access,* North Holland: Amsterdam.

Dalenius, T. and Reiss, S.P. (1982), "Data-swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inference*, **6**, 73-85.

Dandekar, R. A., Domingo-Ferrer, J. and Sebe, F. (2002), "LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York.

Dandekar, R., Cohen, M. and Kirkendal, N. (2002), "Sensitive Microdata Protection Using Latin Hypercube Sampling Technique," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002), "Practical Data-Oriented Microaggregation for

Statistical Disclosure Control," *IEEE Transactions on Knowledge and Data Engineering*, **14** (1), 189-201.

Duncan, G. T., Keller-McNulty, S. A.,and Stokes, S. L. (2001), "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map," Los Alamos National Laboratory Technical Report LA-UR-01-6428.

Fienberg, S. E. (1997), "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, commissioned by Committee on National Statistics of the National Academy of Sciences.

Fienberg, S. E., Makov, E. U. and Sanil, A. P., (1997), "A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data," *Journal of Official Statistics*, **14**, 75-89.

Fienberg, S. E., Makov, E. U. and Steel, R. J. (1998), "Disclosure Limitation using Perturbation and Related Methods for Categorical Data," *Journal of Official Statistics*, **14**, 485-502.

Fuller, W. A. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, **9,** 383-406.

Iyengar, V. (2002), "Transforming Data to Satisfy Privacy Constraints," Association of Computing Machinery, Special Interest Group on Knowledge Discovery and Datamining '02.

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 303-308.

Kim, J. J. (1990), "Subdomain Estimation for the Masked Data," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 456-461.

Kim, J. J. and Winkler, W. E. (1995), "Masking Microdata Files,"American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 114-119.

Little, R. J. A. (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics*, **9,** 407-426.

Little, R. J. A. and Liu, F. (2002), "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, to appear.

Little, R. J. A. and Liu, F. (2003), "Comparison of SMIKe with Data-Swapping and PRAM for Statistical Disclosure Control of Simulated Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, to appear.

Moore, R. (1995), "Controlled Data Swapping Techniques For Masking Public Use Data Sets," U.S. Bureau of the Census, Statistical Research Division Report rr96/04, (available at http://www.census.gov/srd/www/byyear.html).

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., and Sollenberger, P. (1998), "A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models," Survey Research Center, University of Michigan.

Raghunathan, T.E., Reiter, J. P. and Rubin, D.R. (2003), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, **19**, 1-16.

Reiter, J.P. (2002), "Satisfying Disclosure Restrictions with Synthetic Data Sets," *Journal of Official Statistics*, **18**, 531-543.

Reiter, J.P. (2003), "Methods of Inference for Partially Synthetic, Public Use Data Sets," *Journal of Official Statistics*, to appear.

Rubin, D. B. (1993), "Satisfying Confidentiality Constraints through the Use of Synthetic Multiply-imputed Microdata,"*Journal of Official Statistics*, **91**, 461-468.

Samarati, P. (2001), "Protecting Respondents' Identity in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering*, **13** (6), 1010-1027.

Samarati, P. and Sweeney, L. (1998), "Protecting Privacy when Disclosing Information: k-anonymity and its Enforcement through Generalization and Cell Suppression," Technical Report, SRI International.

Sweeney, L. (2002), "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 571-588.

Thibaudeau, Y. and Winkler, W.E. (2002), "Bayesian Networks Representations, Generalized Imputation, and Synthetic Microdata satisfying Analytic Restraints," Statistical Research Division report RRS 2002/09at http://www.census.gov/srd/www/byyear.html.

Trottini, M. and Fienberg, S. E. (2002), "Modelling User Uncertainty for Disclosure Risk and Data Utility," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 511-528.

Winkler, W. E. (1997), "Views on the Production and Use of Confidential Microdata," Statistical Research

Division report RR 97/01 at http://www.census.gov/srd/www/byyear.html.

Winkler, W. E. (2002), “Single Ranking Micro-aggregation and Re-identification,” Statistical Research Division report RRS 2002/08 at http://www.census.gov/srd/www/byyear.html.

Yancey, W.E., Winkler, W.E., and Creecy, R. H. (2002) “Disclosure Risk Assessment in Perturbative Microdata Protection,” in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York (also report RRS 2002/01 at http://www.census.gov/srd/www/byyear.html)

Session 3

The Impact of the Visible:

Images, Spacing, and Other Visual Cues in Web Surveys

The Impact of the Visible:
Images, Spacing, and Other Visual Cues in Web Surveys

Roger Tourangeau, Mick P. Couper, and Fred Conrad
Survey Research Center, University of Michigan
Joint Program in Survey Methodology, University of Maryland

Introduction

The rapid acceptance of the Web as a vehicle for survey data collection raises important questions for survey designers. Web surveys are the latest example of computerized self-administration of survey questions, and we suspect they may ultimately turn out to be the most popular. Aside from the gains from computerization and self-administration, Web data collection eliminates interviewers entirely, sharply reducing the cost of data collection. Furthermore, Web surveys can deliver rich visual content that is impossible or prohibitively expensive to incorporate in other modes. Not surprisingly, the growth in Web surveys has been dramatic. Despite serious concerns about coverage and nonresponse in Web surveys (Couper, 2001), the commercial research sector has rapidly embraced the Internet for faster and cheaper data collection, and almost daily there are reports of new surveys being done over the Web.

A key characteristic of Web surveys is their reliance on visual presentation of the questions. Of course, sound can be added to Web questionnaires, but so far Web surveys have remained a visual medium. Visual presentation is not unique to Web data collection, but is shared to varying degrees with most other methods of self-administration, including mail surveys.

Still, the implications of visual presentation are not especially well understood, even for the older methods; the literature on the design of mail or paper-based self-administered questionnaires is not large. Although several good texts offer practical guidelines for the design of paper self-administered questionnaires (e.g., Dillman, 1978; Mangione, 1995), there has been relatively little empirical work or theoretical analysis of the issues involved. The forms design literature is sparse in general (see, e.g., Burgess, 1984; Waller, 1984; Wright and Barnard, 1975). The one notable exception has been the work of Redline and Dillman, who have applied principles rooted in visual perception theory to the design of self-administered forms (Dillman, Redline, and Carley-Baxter, 1999; Jenkins and Dillman, 1995; Redline and Dillman, 2002). The focus of this work has been on designing forms so that respondents are willing and able to complete them. But the design of paper forms and computer screens may affect not only whether respondents answer the questions but also which answers they give (e.g., Sanchez, 1992; Smith, 1995). The study of forms design is in its infancy, and the impact of forms design on measurement error has been almost entirely neglected.

The studies we present here support a few general conclusions about the impact of visual information on responses to questions in Web surveys:

- Respondents notice images in Web surveys and the content of these images can affect the answers they give;

- Respondents also take in such visual cues as the spacing and relative position of the response options and these cues can alter their interpretation of survey questions;

- Respondents are sensitive to information that is immediately visible and may ignore information that is equally critical but not equally available.

Taken together, our results suggest that, whether we want them to or not, respondents attend to the visual design of Web questionnaires as well as to the verbal content of the questions.

Images as Context

One line of our work has focused on the use of photographic images to supplement question text. As we have argued in an earlier paper (Couper, Tourangeau, and Kenyon, in press), visual and verbal elements may be essential to complete the task of understanding and responding to the questions or these elements may be inessential stylistic embellishments that create an overall "look-and-feel" for the questionnaire. This task-style continuum suggests several different ways pictures can be used in Web surveys:

1. Questions in which images play an essential role (such as questions on recall of an advertisement, brand recognition questions, questions on magazine readership, etc.);

2. Questions in which images supplement the question text, whether the images are intended as motivational embellishments or as illustrations of the meaning of the question;

3. Questions in which the images are incidental (providing branding, an attractive background, etc.).

All three combinations of text and image appear to be quite widespread in Web surveys. The arguments for questions using the first type of text-image combination are quite compelling, and questions in the third category — in which the images are incidental to the task — may also make sense in the highly competitive world of Web surveys, where branding is an important goal of many purveyors of Web surveys and services. Questions in which images are intended to play a supplementary role are potentially the most problematic, because it may not be clear to respondents whether the images are intended as task elements or style elements.

Whether the survey designers intend it or not, images can serve as powerful contextual cues that alter what material comes to mind as respondents formulate their answers; they can affect how respondents construe the targets of their judgments or the standards they apply in making those judgments. Let us briefly summarize the results of three studies that illustrate these processes.

Images, target categories, and frequency judgments. Our first experiments on the impact of images were done by Knowledge Networks, which embedded them in a survey administered to a sample of U.S. adults from the Knowledge Networks panel. The panel is made up of approximately

100,000 panel members from almost 50,000 households in the United States, initially recruited from a list-assisted RDD sample. Each panel member receives the same WebTV unit and software, which help assure that the survey looks the same to every panel member. Some 56% of contacted households agree to join the panel but only 80% of those actually install the WebTV unit and only 83% of those complete the initial questionnaire which gathers basic demographic data on panel members. These are average estimates, as panel recruitment is an ongoing effort. Dennis (2001) provides more details on the design and implementation of the panel. About 3,000 members of the panel were asked to complete our survey, and 2,385 of them did. Taking into account the losses at earlier stages of recruitment and data collection, the cumulative response rate for our survey was 30%.

The survey, which concerned travel, leisure, and shopping activities, included six parallel experiments summarized in Table 1 below. All six followed the same logic. For each topic, we developed four versions of the questions:

1. a version that did not include any picture (the *no picture* condition);

2. a version featuring an image of a salient, but low frequency instance of the behavior in question (the *low frequency* condition);

3. a version featuring an image of a salient high frequency instance (the *high frequency* condition); and

4. a version that displayed both pictures (the *both pictures* condition).

Our hypothesis was that presenting the picture of the high frequency instance would enhance the retrieval of similar instances and increase the total number of instances reported. By contrast, the picture of the low frequency instance would trigger the recall of relatively infrequent incidents similar to the one in the picture. For example, we asked respondents about their shopping trips in the past month and expected that showing them a picture of a grocery store would increase the overall number of shopping trips they reported on average compared to the picture of a department store, since trips to the grocery (cued by the one picture) are likely to be more frequent than trips to a department store (cued by the other).[1]

[1] For two of the topics in our study, we carried out a follow-up study to confirm that the pictures did in fact portray highly salient instances of the category. The follow-up questionnaire included questions asking the respondents how often they went shopping and how often they took overnight trips. Just after the frequency question on shopping, respondents were asked "which of the following types of store did you consider in answering the previous question," with grocery stores and department stores among the possibilities listed. (Respondents were asked to pick all of the types of store they had considered.) Similarly, we asked respondents "which of the following types of trips" they had in mind in answering the prior question on their travel frequency. Grocery stores were the most commonly mentioned type of store, with 93.2% of the respondents indicating they had considered them in responding to the item about how often they went shopping. Department stores were the next most popular choice (64.9%; another 5.9% mentioned clothing stores but not department stores). For the travel item, the most popular choices were family vacations by car (76.9%), family visits by car (65.6%), and vacations by plane (50.2%). Business trips by plane were

Table 1. Images Displayed (and Sample Sizes) in Study 1, by Condition and Topic

Question topics	Picture Descriptions			
	No Picture	Low Frequency Instance	High Frequency Instance	Both Pictures
Overnight trips in last year	(579)	Businessman at airport (620)	Family station wagon (593)	(593)
Sporting events attended in last year	(582)	Large baseball stadium (621)	Little league ball game (646)	(536)
Times went out to eat in past month	(592)	Intimate restaurant (593)	Eating fast food in a car (585)	(615)
Live music events attended in the last year	(608)	Large outdoor rock concert (608)	Piano and singer at club (572)	(597)
Listening to recorded music in the past week	(591)	Listening to the hi-fi (588)	Listening to the car radio (598)	(608)
Shopping trips in the past month	(616)	Department store (clothing) (594)	Grocery store (548)	(627)

We compared the four means for each topic using one-way ANOVAs. For all six topics, the overall *F*-tests were significant. In addition, for four of the six topics, the means for the high and low frequency conditions differed significantly from each other (at $p < .01$ or less); the two exceptions involved live music and recorded music. In all four cases, the difference was in the expected direction, with the pictures showing the high frequency instances of the behaviors prompting higher reporting on the average than the pictures showing the low frequency instances. We interpret this as the same sort of accessibility-based context effects that are often found in attitude surveys (see Chapter 7 in Tourangeau, Rips, and Rasinski, 2000) — the images affect the number and type of instances respondents retrieve in formulating their answers ("priming" those memories); the number and type of instances retrieved in turn affect the judged frequency of the behavior.

Responses to an open-ended debriefing question at the end of our second survey suggested that the pictures may not only have primed specific memories, but also affected how respondents construed the category of interest. This was most noticeable for the question on shopping frequency, which

mentioned by 24.9% of the respondents. In the absence of any pictures, then, respondents were likely to consider these instances in assessing the frequency of shopping and traveling — they are highly salient examples. Still, for some respondents, the pictures were likely to remind them of incidents they might otherwise have forgotten or overlooked.

was followed by a question on the proportion of shopping trips that were for food. Several respondents commented on the impact of the images, for example:

> "What kind of shopping you were looking for was not defined because my number of times would be different depending on what type. I took it as how many times for leisure." [No picture]
>
> "Thought shopping meant clothes from picture. If you include food shopping — went about 10 times" [Department store picture]
>
> "I shop for groceries almost every week. Does that count? The pictures are nice, but add to the time it takes to answer a survey." [Department store picture]
>
> "The pictures helped remind me that a little league game is just as much a sporting event as a trip to Fenway. The pics were a help." [Both sporting event pictures]

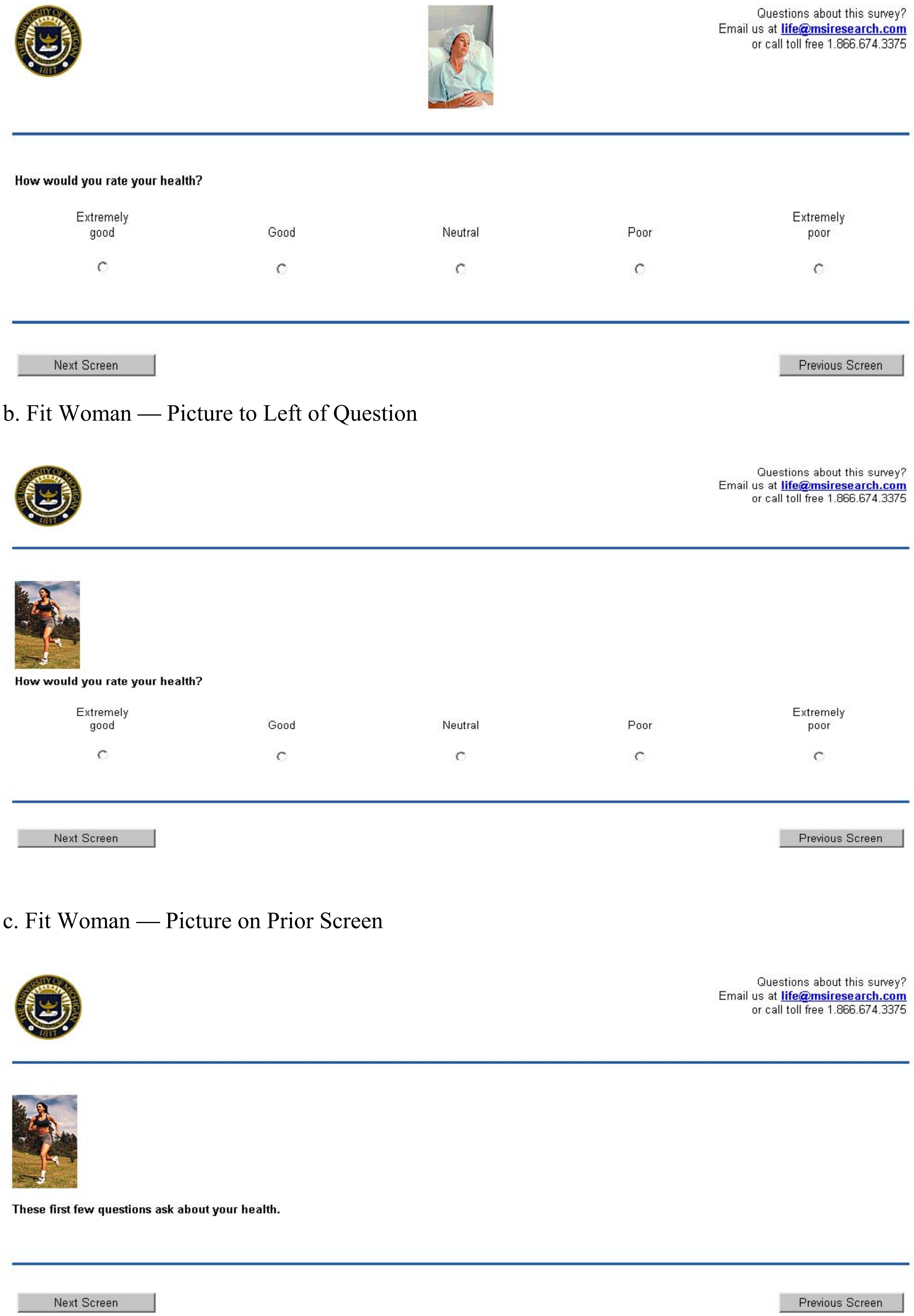
For some respondents, the pictures clarified the meaning of the questions, broadening their definition of the target category. For others, the pictures may have reinforced a relatively narrow interpretation of the question's meaning.

Images and rated health. In our initial studies, then, respondents exhibited what are sometimes called *assimilation* effects in the context effects literature. When they saw images of high frequency events, they reported higher frequencies; when they saw images of low frequency events, they reported lower frequencies. Verbal context (in the form of prior items) can sometimes have the opposite effect on answers to subsequent questions. When the prior questions suggest an extreme standard of comparison that respondents apply in judging later items, the target judgments are pushed in the direction opposite of the standard. For example, respondents may report liking a politician less when they rate an extremely popular politician first (Schwarz and Bless, 1992). We thought we could create similar judgmental contrast effects using images rather than prior questions to set the standard for the target items.

This experiment was embedded in a Web survey conducted by MSInteractive. In March and April of this year, MSInteractive sent e-mail invitations to 39,217 members of SSI's Web survey frame. The e-mail invitation asked them to complete a survey of attitudes and lifestyles sponsored by the National Science Foundation; it included the URL for the questionnaire. The SSI frame consists of some seven million e-mail addresses collected at various Web sites. A total of 3,179 persons started the questionnaire, 2,722 of them getting all the way through it. The response rate was 6.9 percent (not counting the partials) or 8.1 percent (counting them).

The experiment compared the impact of two pictures on respondents' judgments of their overall health (that is, responses to an item asking, "How would you rate your health?"). One group of respondents saw photograph of a healthy young woman jogging; another group saw a picture of a woman in a hospital bed. The experiment also compared three different positions for the picture — on the prior screen just before the health item, on the same screen in the survey header, or just to the left of the question text. Figure 1 displays examples of the pictures we used.

Figure 1. Images used in Study 2



a. Sick Woman — Picture in Header

Questions about this survey?
Email us at **life@msiresearch.com**
or call toll free 1.866.674.3375

How would you rate your health?

Extremely good	Good	Neutral	Poor	Extremely poor
○	○	○	○	○

b. Fit Woman — Picture to Left of Question

Questions about this survey?
Email us at **life@msiresearch.com**
or call toll free 1.866.674.3375

How would you rate your health?

Extremely good	Good	Neutral	Poor	Extremely poor
○	○	○	○	○

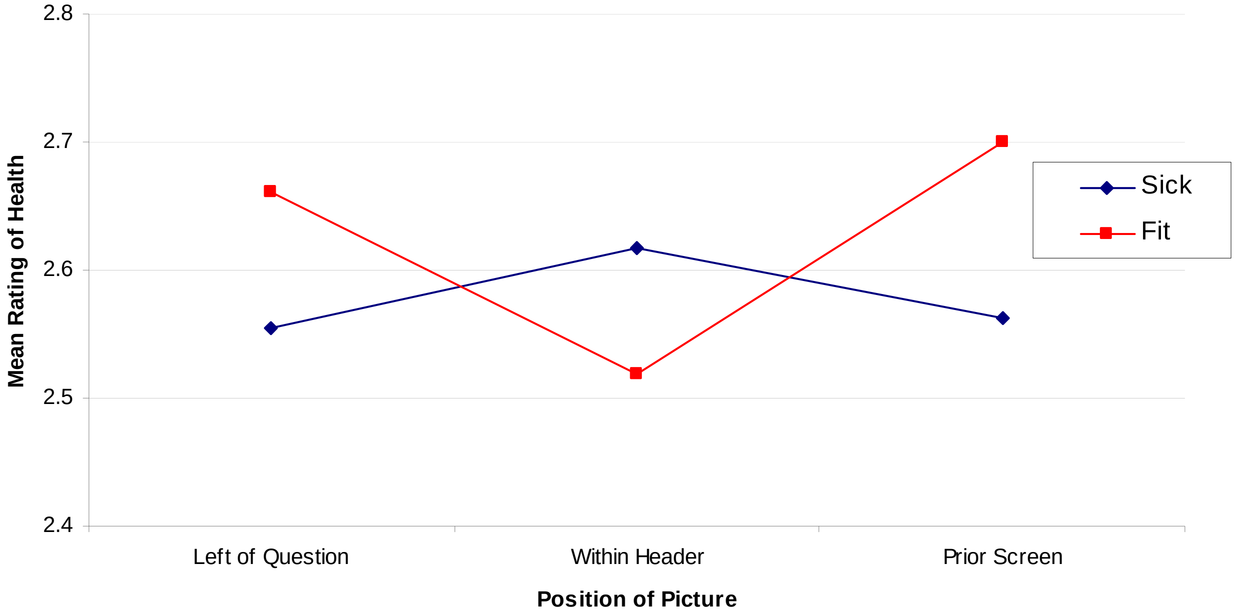
c. Fit Woman — Picture on Prior Screen

Questions about this survey?
Email us at **life@msiresearch.com**
or call toll free 1.866.674.3375

These first few questions ask about your health.

As expected, the pictures affected the self-ratings of health, lowering them on average for the respondents who got the picture of the healthy woman jogging (mean of 2.64) and raising them on average for those who got the picture of the sick woman in bed (2.58). (Higher numbers indicate worse health.) The overall effect of the picture was only marginally significant — $F(1, 2309) = 3.08, p < .08$. But we didn't expect the significant interaction between the position of the picture and its content; that interaction is displayed in Figure 2. When the picture is in the in the header, assimilation rather than contrast seems to be the result.

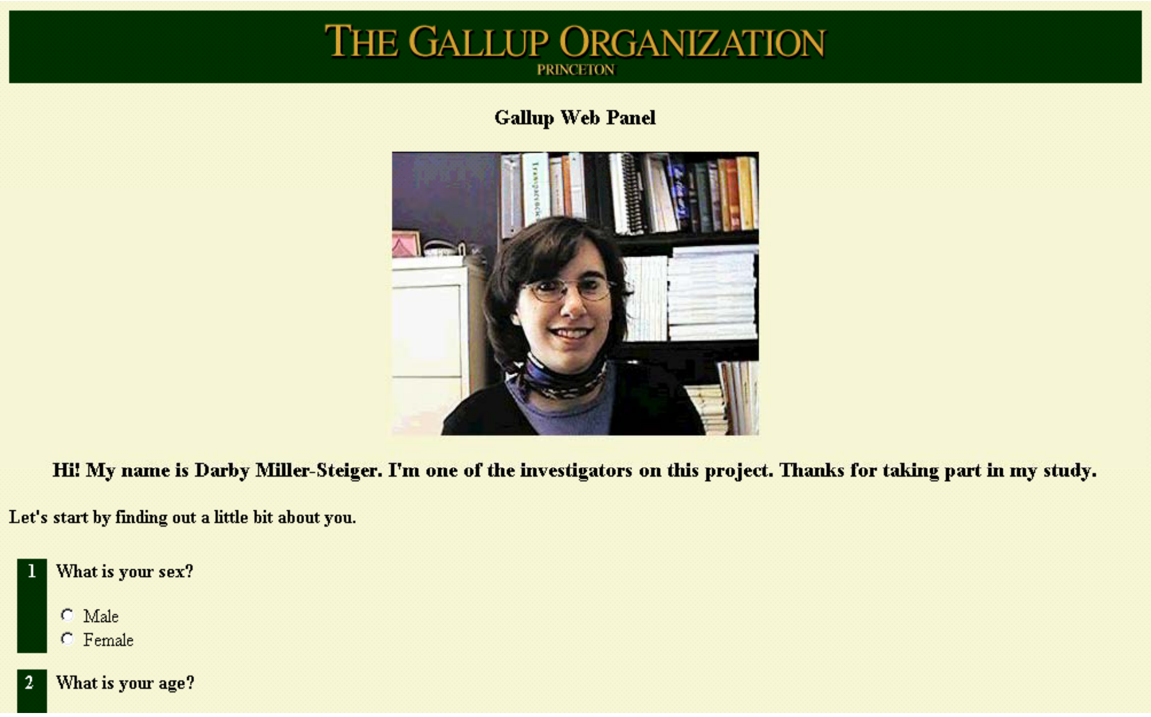
Figure 2. Health Ratings, by Picture and Position



At least in some conditions, then, images provide a standard of comparison against which our judgments of later targets, such as our own health, are contrasted.

Images in the interface. Tourangeau, Couper, and Steiger (2003) reported another series of experiments that incorporated images as part of the interface of a Web survey. (These studies were done by the Gallup Organization. Because these studies have been published, we omit the methodological details here.) Figure 3 below shows an example of the images the interface incorporated. The opening screen displayed the female face of one of the investigators (Steiger); other versions of the questionnaire displayed a male picture. Across two separate Web surveys, we examined the impact of the interface on answers to a variety of questions. For the most part, it didn't matter whether the survey had a male or a female "face," but for one set of items it did. These were a battery of questions on sex roles that are known to be affected by the sex of the (live) interviewer (Kane and Macauley, 1993). Men and women both give more pro-feminist responses to these items when female interviewers administer them than when male interviewers do. We found a similar pattern with our "virtual" interviewers; the responses were more pro-feminist when the survey had a female “face” (as in Figure 3) than a male one. We suspect that this is a priming effect; when the respondents see the picture of an attractive working woman, it tends to bring to mind consistent (that is to say, positive) thoughts about women in the work place. The male interface tends to bring to mind more traditional views about the roles of men and women.

Figure 3. “Female” Interface used in Web Surveys



THE GALLUP ORGANIZATION
PRINCETON

Gallup Web Panel

Hi! My name is Darby Miller-Steiger. I'm one of the investigators on this project. Thanks for taking part in my study.

Let's start by finding out a little bit about you.

1 What is your sex?

○ Male
○ Female

2 What is your age?

Researchers in human-computer interaction tradition have reported even more striking results. For example, Walker, Sproull, and Subramani (1994) administered questionnaires to people using either a text display or one of two talking-face displays to ask the questions. Those interacting with a talking-face display spent more time, made fewer mistakes, and wrote more comments than did people interacting with the text display. However, people who interacted with an expressive face liked the face and the experience less than those who interacted with an inexpressive face. In another experiment, Sproull and colleagues (1996) varied the expression of a talking face on a computer-administered career counseling interview; one face was stern, the other pleasant. The faces were computer-generated images with animated mouths. They found that:

> People respond to a talking-face display differently than to a text display. They attribute some personality attributes to the faces differently than to a text display. They report themselves to be more aroused (less relaxed, less confident). They present themselves in a more positive light to the talking-face displays. (p. 116).

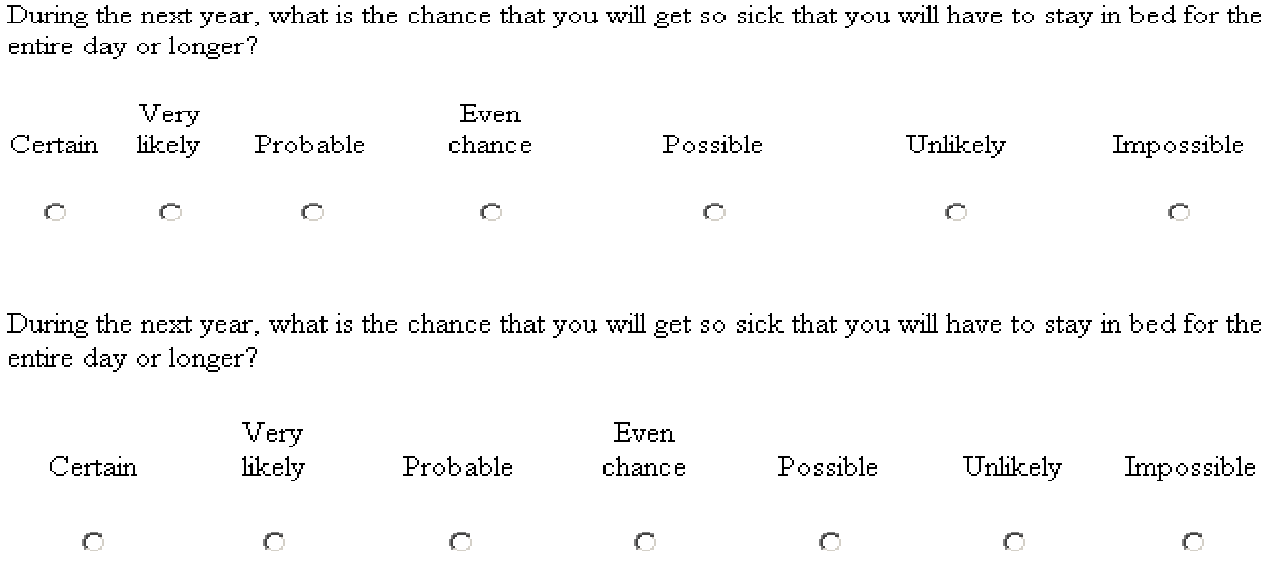
The interface to a survey, particularly when it incorporates humanizing visual cues, may itself constitute a contextual stimulus, one that is capable of altering respondents' views of the survey and their responses to the questions.

Spacing and Position

Spacing of the response options. Our experiment on self-rated health investigated a second issue besides the effect of the photographs. The question following the standard health item asked respondents how likely it was they'd get sick enough during the next year that they have to spend a day or more in bed ("During the next year, what is the chance that you will get so sick that you will have to stay in bed for the entire day or longer?"). We varied the spacing of the response options that made up the scale on which respondents were to indicate their answers. This experiment is one of a number we've done that share the notion that respondents follow simple heuristics in interpreting the visual features of questions. Though these interpretive heuristics are often useful, they may sometimes lead to unintended inferences about the meaning of a question. Hoffman (2000) argues that interpretive rules are central in visual processing and are responsible for such key abilities as depth perception. The heuristics for interpreting visual stimuli can sometimes lead to systematic misinterpretations of those stimuli, producing optical illusions. In the same way, the application of interpretive heuristics for visual cues in questionnaires can lead to erroneous inferences about the meaning of survey questions.

One of these heuristics involves seeing the option that is physically in the middle of the scale as representing the scale midpoint; we refer to this as the "middle means typical" heuristic. We varied the spacing of the response options to the question about the chance of a sick day in bed. Approximately half of the respondents got the item with evenly spaced response options (see Figure 4); the remainder got a scale in which four of the seven options were to the left of the visual midpoint of the scale.

Figure 4. Scales Used in Experiment on Spacing of Response Options



During the next year, what is the chance that you will get so sick that you will have to stay in bed for the entire day or longer?

Certain ○ Very likely ○ Probable ○ Even chance ○ Possible ○ Unlikely ○ Impossible ○

During the next year, what is the chance that you will get so sick that you will have to stay in bed for the entire day or longer?

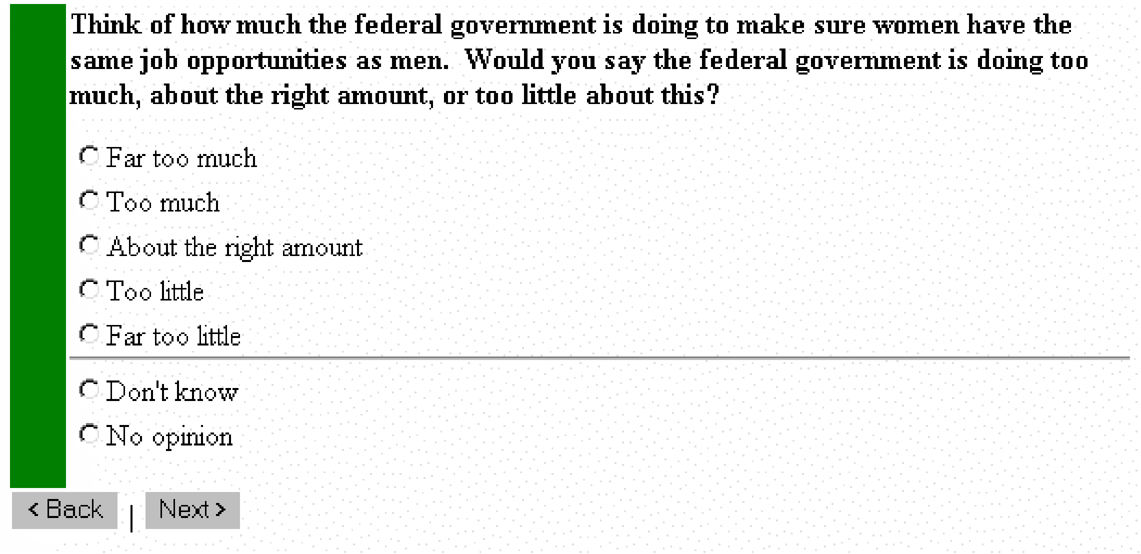
Certain ○ Very likely ○ Probable ○ Even chance ○ Possible ○ Unlikely ○ Impossible ○

The ratings were significantly higher when respondents got the unevenly spaced scale (the top one in Figure 4) rather than scale that arrayed the response categories evenly (the bottom one). The means were 4.60 in the even spacing condition versus 4.45 in the uneven spacing condition; *F*(1, 3083) = 7.58, *p*< .01.

Separating substantive and nonsubstantive options. In one of our Gallup surveys, we did another study that demonstrated the importance of the spacing of the response options. That experiment compared two methods of separating nonsubstantive response options (Don't know, Refused) from substantive ones. In one case, the nonsubstantive options were simply presented as additional radio buttons; in the other, we included a divider line that clearly separated the nonsubstantive options from the rest (see Figure 5).

Figure 5. Formats for Displaying Nonsubstative Options

a. Divider Line Version

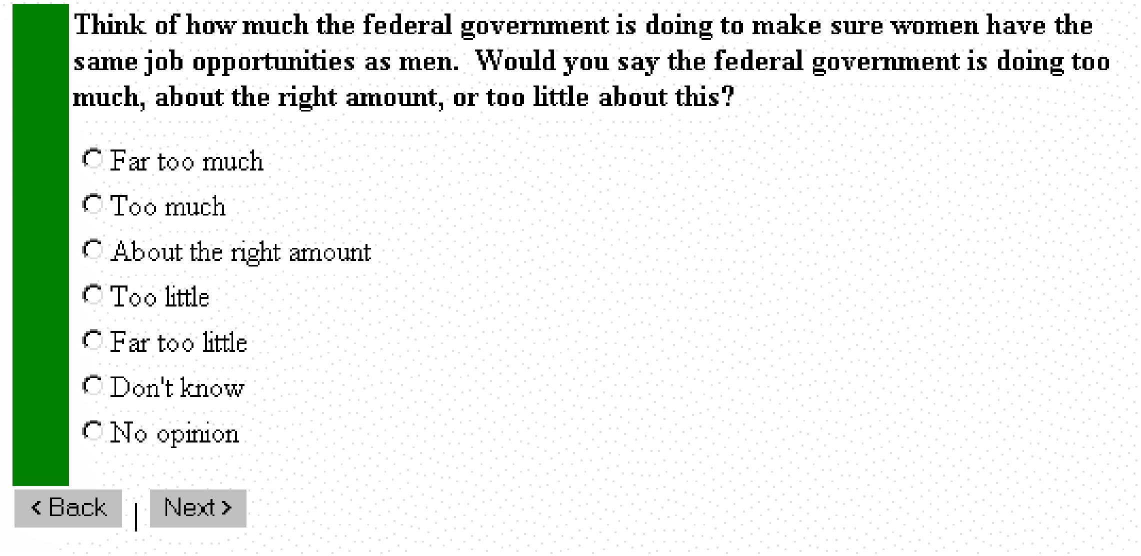


Think of how much the federal government is doing to make sure women have the same job opportunities as men. Would you say the federal government is doing too much, about the right amount, or too little about this?

○ Far too much
○ Too much
○ About the right amount
○ Too little
○ Far too little

○ Don't know
○ No opinion

b. Version with No Divider Line



Think of how much the federal government is doing to make sure women have the same job opportunities as men. Would you say the federal government is doing too much, about the right amount, or too little about this?

○ Far too much
○ Too much
○ About the right amount
○ Too little
○ Far too little
○ Don't know
○ No opinion

In the version with the divider line, the visual midpoint of the scale falls at the conceptual midpoint ("About the right amount"). In the version without the divider, the visual midpoint actually falls on one end of the scale ("Too little"). This difference affected the average responses — there's a significantly lower mean without the divider than there is with it. Moreover, the divider line seemed to draw attention to the nonsubstantive options; there are significantly more nonsubstantive answers given when the divider line is displayed (21.4%) than when it's omitted (17.5%).

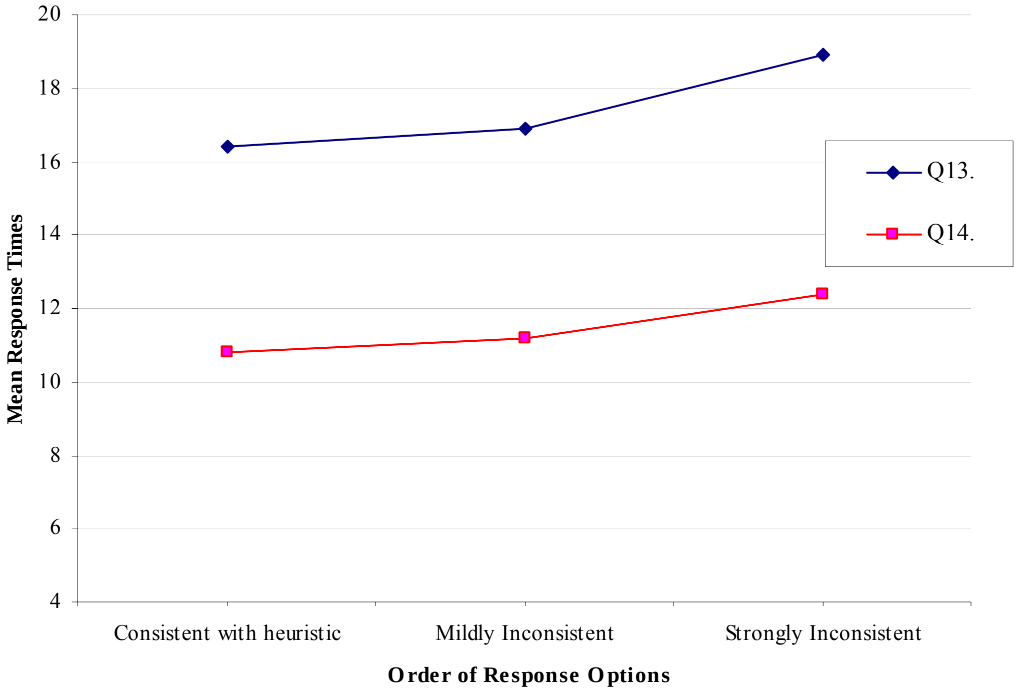
Positional inferences. Another heuristic respondents may use in understanding and applying response options is the "Left and top mean first" heuristic. According to the heuristic, the leftmost or top item in a list of items represents the "first" in some conceptual sense. For example, when the list is a series of ordered response categories or scale values, respondents expect the top or leftmost option to represent one of the two endpoints ("Agree strongly") and they expect each of the successive options to follow in some logical order ("Agree," "Neither agree nor disagree," and so on). If the list does not conform to these expectations, respondents may become confused, make mistakes, or take longer to respond.

Our first Web study with MSInteractive experimentally varied the order of the response options in six of the survey questions.2 We carried out two independent experiments, one with four frequency items and the other with two agree-disagree items. We focus here on the results from the agree-disagree items. Each experiment compared three versions of the questions. In one version, the response options followed the logical order. For the agree-disagree items, this version went from "Strongly agree" to "Strongly disagree" with "It depends" in the middle. A second version presented the options in order of decreasing agreement, with "It depends" as the final option). In the final version, "It depends" was the first option presented, but the remaining options were ordered by extremity ("Agree strongly," "Disagree strongly," "Agree," and "Disagree"). Respondents got all the response options in the same order for all four of the frequency items; similarly, they randomly assigned to receive one of the three versions for both agree-disagree items.

We anticipated that respondents would answer the questions most quickly when the items followed the order implied by the "left is first" heuristic, with the slowest answers in the third version of the questions (where the order of the response categories departs most sharply from the order implied by the heuristic). Three of the six items showed significant differences in response times and all three show the expected pattern. Figure 6 below displays the average response times for the two agree-disagree items — Q13 ("It is SENSIBLE to do exactly what the doctors say") and Q14 ("I have to be VERY ILL before I go to the doctor") in the experiment. For both items the differences in reaction times across experimental treatments were highly significant: *F*(2,2533)=18.7 for Q13 and *F*(2,2591)=12.6 for Q14.

2In February and March of 2002, MSInteractive conducted a Web survey, in which 14,192 e-mail invitations were sent to members of SSI's Web survey frame. The e-mail invitation asked them to complete a survey sponsored by the National Science Foundation. A total of 2,871 persons started the questionnaire, 2,568 of them getting all the way through it, for a response rate of 18.1% (not counting the partials) or 20.2% (counting them). Among other experiments, the survey included one that compared three response formats.

Figure 6. Response Times and Consistency with Heuristic



Another implication of the "Left is first" heuristic is that respondents may use it to infer the characteristics of an unfamiliar item from its position in a list of similar items. For example, we compared the percentage of respondents rating the Fiat Tipo as an expensive car when it came third in a list of cars that included the BMW 318, Acura Integra, Mazda Protégé, Toyota Corolla, Dodge Neon, and Geo Metro to when it came last seventh — right after the Geo Metro. Respondents were significantly more likely to say the Tipo was an expensive car when it came third in the list (72.4%) than when it came last (60.3%; $\chi^2 = 45.3$, df =1, $p < .001$). We found similar results for three out of five other items (see Table 2 below).

Table 2. Proportion Yes (and Sample Size), by Position in List

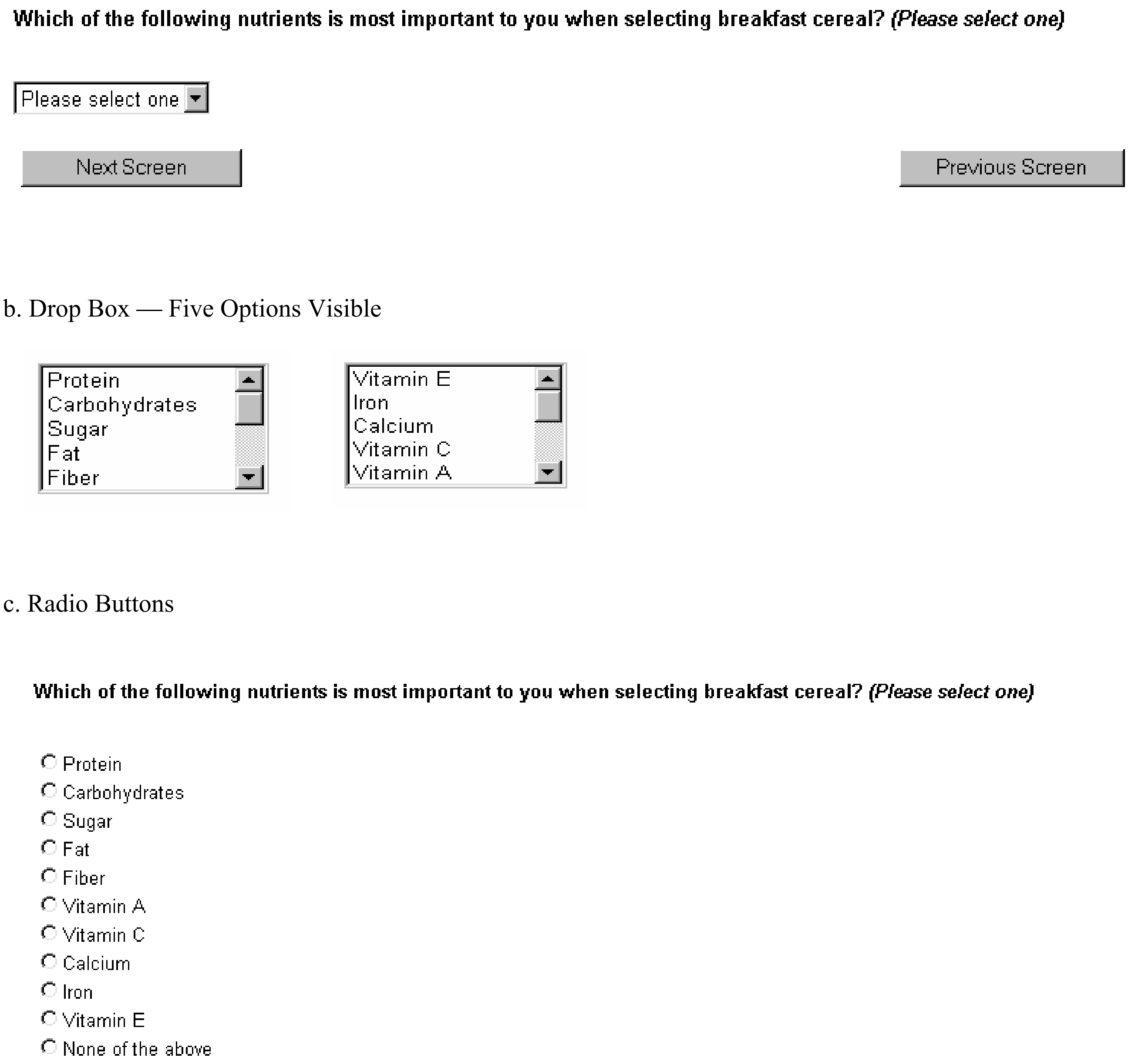
Judgment/Item	**Percent Yes** (n)	
	Third in List	Seventh in List
Important for healthy diet/Isoflavin	44.4 (1396)	43.2 (1326)
Low in saturated fat/Cod liver oil	42.7 (1396)	36.7 (1326)
Expensive hotel/Clarion Inn	61.5 (1396)	44.3 (1326)
Expensive city/Ocala,FL	51.5 (1396)	59.1 (1326)
Expensive midsize/Austin Rover	92.0 (1396)	86.0 (1326)
Expensive small car/Fiat Tipo	72.4 (1396)	60.3 (1326)

Note: The differences between columns are significant (p < .05 or less) for all but the first row.

The Visibility Principle

Images, spacing, and positioning all have an impact on the answers. We have also looked at how the method of presenting the response options can affect the distribution of the answers. We compared radio buttons (a format that displayed all 11 answer options from the outset), drop down boxes in which only the first five answer options were visible initially, and drop down boxes in which none of the options were visible until respondents clicked on the drop down arrow. Figure 7 shows the key conditions. In addition, the experiment varied the order of the answer options. Approximately half of the respondents got the response options in one order; the rest got them in the reverse order.

Figure 7. Formats Compared in Response Format Study



a. Drop Box — None of the Options Visible

Which of the following nutrients is most important to you when selecting breakfast cereal? *(Please select one)*

Please select one ▾

Next Screen

Previous Screen

b. Drop Box — Five Options Visible

Protein
Carbohydrates
Sugar
Fat
Fiber

Vitamin E
Iron
Calcium
Vitamin C
Vitamin A

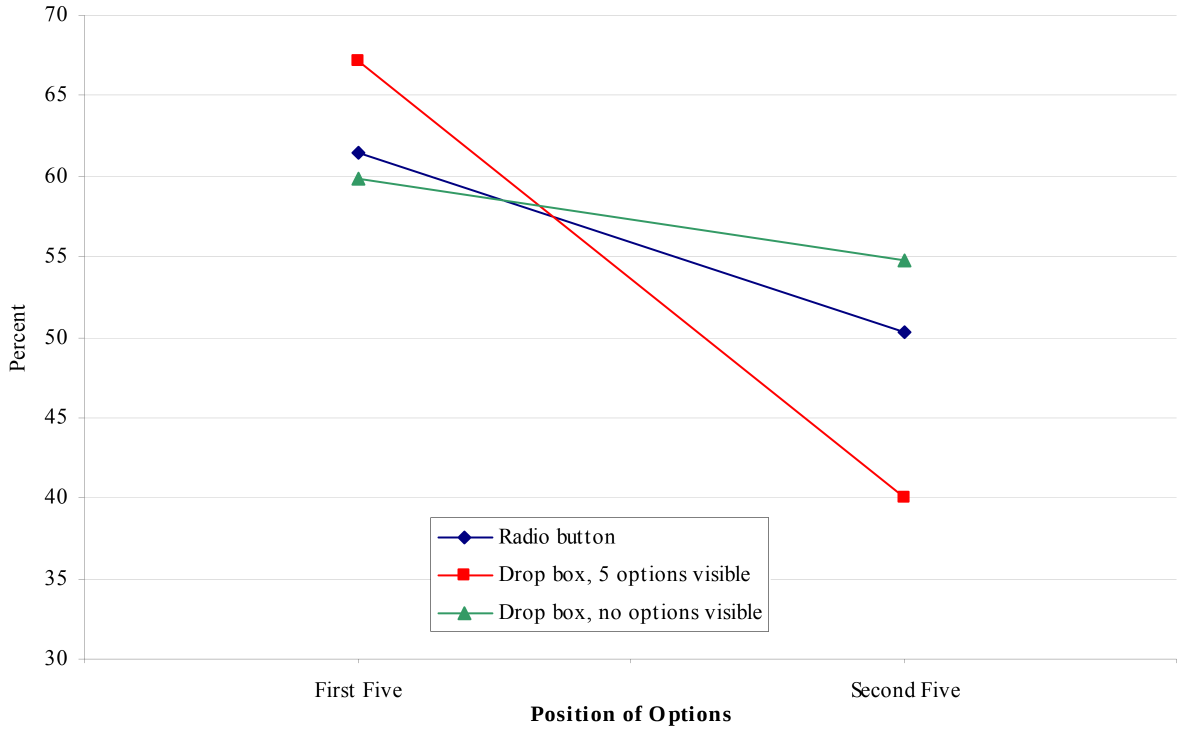
c. Radio Buttons

Which of the following nutrients is most important to you when selecting breakfast cereal? *(Please select one)*

○ Protein
○ Carbohydrates
○ Sugar
○ Fat
○ Fiber
○ Vitamin A
○ Vitamin C
○ Calcium
○ Iron
○ Vitamin E
○ None of the above

Our main hypothesis was the respondents would focus on the options they could see and thus would be more likely to select one of the five options displayed initially in the drop down box. Figure 8 displays the key result. In all three response formats, respondents were more likely to select one of the first five options listed in Figure 7c. above ("Protein" through "Fiber") when these were the first five options listed than when they were the final five. This is a classic primacy effect and similar effects are often found with items displayed visually (Krosnick & Alwin, 1987; Krosnick, 1991); however, the effect is far more marked in the drop down box condition in which only the first five options were visible. We included a replication of this experiment (Q20) later in the questionnaire, and the results are similar to those in Figure 8. The key interaction of response order and response format was highly significant for both items. When it takes additional effort to see some of the options, respondents are especially unlikely to choose them. Redline and Dillman (2002; see also Jenkins and Dillman, 1995) have also emphasized the importance of visual prominence.

Figure 8. Impact of Response Format and Response Order



Getting help. Our first MSInteractive survey also included experiments that examined the impact of the accessibility and usefulness of on-line definitions for survey terms. Respondents were asked to evaluate (on a five-point scale) whether they consume as much as they should of four food/nutrition products. They were told that they could obtain a definition for any of the terms by clicking on them but were not specifically instructed to do so. The primary concern in the study was how often they obtained definitions. Our initial hypotheses were that respondents would be deterred from obtaining definitions if it was hard to access them or if the definitions did not provide useful information, information relevant to the respondent's judgment.

Respondents were able to obtain definitions in one of three ways that varied the number of clicks required: (1) they could display the definition by simply clicking on the highlighted term in the question (one click); (2) they could display the definition by first clicking on the highlighted term, which displayed a list of all terms for which definitions were available, and then by selecting (clicking on) the term of interest in the list (two clicks); or (3) they could display a definition by clicking at least twice, first on the highlighted term which displayed a text file glossary, and then by scrolling the glossary (clicking the scroll bar at least once) to locate the definition of interest. We created "useful" definitions by including some surprising information that might alter respondents' judgments, e.g., the fact that vegetables include French fries. The idea was that respondents might answer differently when they read this kind of definition than when they did not. In contrast, the definitions we created to be "not useful" presented information that was unlikely to affect respondents' answers. Consider the definition for hydrogenated fat: "A fat that has been chemically altered by the addition of hydrogen. Vegetable shortening and margarine are hydrogenated fats." The information in the definition is accurate but not very helpful in evaluating one's consumption of hydrogenated fat.

It is probably not a surprise that respondents tended to ignore the definitions when they had to do something to make them visible. Only 17.4% of the respondents (a total of 501 of them) obtained definitions. This is quite low considering that definitions may, potentially, be essential for respondents to interpret questions in the intended way. The respondents may have been unaware that question terms could have special meanings or the instructions may not have indicated the potential value of the definitions. When respondents did obtain at least one definition, they did so overwhelmingly (89% of the time) for technical terms (e.g. "antioxidants"), where meaning was an obvious concern. Their relatively infrequent requests (11% of the time) for definitions of non-technical terms (e.g. "dairy products") suggest it is easy for respondents to overlook possible differences between their interpretation and the intended one (for example, the definition of dairy products included "cheesy foods like pizza" though many respondents probably would not ordinarily include these). The difference due to the type of terms ($p < .001$) suggests that at least sometimes respondents did not get any definitions because they did not realize they might need them.

Another factor in the low percentage of respondents who accessed definitions was that the amount of effort required; even one click was more than what respondents were willing to expend. Those respondents who did obtain definitions did so far more often when one click was required (56% of the time) than when two or three were required (24% and 20% of the time respectively). The difference due to effort (number of clicks) required ($p < .001$) may indicate that those respondents who never obtained definitions at all were unwilling to invest even one click. The general implication is that interactive features in Web surveys should be designed so that they are very easy to use, requiring no more than one click. When the process involves multiple steps, respondents may begin to invoke the feature, but they are relatively unlikely to complete it. Respondents using the two-click interface started the process by clicking on the highlighted term in the question 629 times but completed it by selecting the term from the list only 246 times. Unless an item is easily seen or it's on the critical path for completing the task, it is unlikely to have much impact on respondents.

Conclusion

Although we've presented quite a few results, we can boil them down to three main themes:

- Respondents attend to pictures and the pictures they see can affect their answers;
- They also attend to the position and spacing of the response options and they use simple heuristics to interpret these and similar visual cues;
- They tend to attend to what's immediately visible and to overlook information they have to make visible.

Consider first the effects of pictures in Web questionnaires. We argue that photographs and other images are powerful contextual stimuli. They can render some instances of a target category more accessible to retrieval than others, leading to assimilation effects. For example, when the pictures display infrequent instances of a category, respondents give lower frequency estimates for the category. By contrast, when the pictures cue more frequent instances, respondents give higher frequency estimates. Similarly, pictures can serve as standards of comparison. Show the respondents a picture of a healthy young woman jogging and they may lower their ratings of their own health. Show them a woman in a hospital bed and it boosts the ratings of their own health. And the effects of pictures are not confined to pictures incorporated into the questionnaire itself, but may extend to photographs intended to create a "human" interface for the survey. The image of a woman professional may subtly alter responses to questions about sex roles.

Our studies also demonstrate the impact of spacing and positional cues on answers. The importance of these and similar cues has been demonstrated repeatedly in the work of Redline and Dillman as well. Like Redline and Dillman (Jenkins & Dillman, 1995; Redline & Dillman, 2002), we find support for the general conclusion that respondents have expectations about the visible aspects of survey questions. They expect a series of items or response options to follow a logical progression from left to right or from top to bottom. They slow down when this expectation is violated. They may infer something about unfamiliar items, such as hotel chains or cars, from their position in a list of similar items. When items are grouped, respondents expect them to be related to each other; as a result, presenting a battery of items on a single screen leads to higher intercorrelations among them then presenting them individually on successive screens (Couper, Traugott, and Lamias, 2001). Finally, respondents expect that the conceptual midpoint of the scale to fall at the visual midpoint. When the visual and conceptual midpoints don't coincide, it throws them off and may affect their answers.

People attend to the information that they see. They give more weight to information that they can see than to information that's not immediately visible. If respondents have to work to see a response option (for example, when a drop box doesn't display all the options initially), they are less likely to select it. If they have to click to see a definition for a key term, they are unlikely to do so; and if they have to click more than once, they are even less likely to bother. It is probably useful to think of visibility as a continuum, ranging from information that can only be seen with great difficulty to information we can hardly ignore. If we want respondents to attend to something, we need to make it not just visible, but visually prominent. Otherwise, they are likely to ignore that information in favor of information that easily seen.

Acknowledgments: The research described here was supported by an NSF grant (SES 0106222) to Roger Tourangeau, Mick Couper, and Fred Conrad. We gratefully acknowledge NSF's support. (Other contributors for the original NSF award include the Department of Energy, Internal Revenue Service, Social Security Administration, National Agricultural Statistics Service, Bureau of the Census, Department of Transportation, Bureau of Labor Statistics, and Office of Justice Programs. The National Center for Education Statistics contributed funds for a supplemental award.) Much of the work carried out here was conducted in collaboration with staff at MSInteractive. We appreciate the contributions of Reg Baker, Scott Crawford, and Duston Pope at MSIinteractive both to the design and implementation of these studies. We also acknowledge the able assistance of Andrey Peytchev.

References

Burgess, J.H. (1984), *Human factors in forms design*. Chicago: Nelson-Hall.

Couper, M.P. (2001). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464-494.

Couper, M.P., Traugott, M., and Lamias, M. (2001). Web survey design and administration. *Public Opinion Quarterly, 65*, 230-253.

Couper, M.P., Tourangeau, R., & Kenyon, K. (In press). Picture this! Exploring visual effects in Web surveys. *Public Opinion Quarterly*.

Dennis, J.M. (2001). Response timing and coverage of non-Internet households: Data quality in an Internet-enabled panel. Paper resented at the annual meeting of the American Association for Public Opinion Research, Montreal, Canada, May.

Dillman, D.A. (1978), *Mail and Telephone Surveys; The Total Design Method*. New York: Wiley.

Dillman, D.A., Redline, C.D., and Carley-Baxter, L.R. (1999). Influence of type of question on skip pattern compliance in self-administered questionnaires. *Proceedings of the American Statistical Association, Survey Research Methods Section*.

Hoffman, D. (2000), *Visual intelligence: How we create what we see*. New York: W. W. Norton and Company.

Jenkins, C.R. and Dillman, D.A. (1995). Towards a theory of self-administered questionnaire design. In L. Lyberg , P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality*. New York: Wiley.

Kane, E.W., & Macauley, L.J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly*, *57*, 1-28.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Krosnick, J., and Alwin, D. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, *52*, 526-538.

Mangione, T.W. (1995). *Mail surveys: Improving the quality*. Thousand Oaks, CA: Sage.

Redline, C.D., and Dillman, D.A. (2002), The influence of alternative visual designs on respondents' performance with branching instructions in self-administered questionnaires. In R. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (Eds.), *Survey nonresponse* (pp. 179-193). New York: John Wiley

Sanchez, M.E. (1992). Effect of questionnaire design on the quality of survey data. *Public Opinion Quarterly*, *56*, 206-217.

Schwarz, N., and Bless, H. (1992). Scandals and public trust in politicians: Assimilation and contrast effects. *Personality and Social Psychology Bulletin*, *18*, 574-579.

Smith, T.W. (1995), Little things matter; a sampler of how differences in questionnaire format can affect survey responses. *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 1046-1051.

Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., and Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, *11*, 97-124.

Tourangeau, R., Couper, M.P., and Steiger, D.M. (2003). Humanizing self-administered surveys: Experiments on social presence in Web and IVR surveys. *Computers in Human Behavior*, *19*, 1-24.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). *The Psychology of survey response*. Cambridge University Press.

Walker, J., Sproull, L., & Subramani, R. (1994). Using a human face in an interface. *Proceedings of the conference on human factors in computers* (pp. 85-99). Boston: ACM Press.

Waller, R. (1984). Designing government forms: A case study. *Information Design Journal*, *4*, 36-57.

Wright, P. and Barnard, P. (1975). "Just Fill in this Form" — A review for designers. *Applied Ergonomics*, *6*, 213-220.

Discussion of
The Impact of the Visible

Cleo Redline
National Science Foundation

I am very honored to be here today to discuss the work presented by Tourangeau and his colleagues. Before I talk about the specifics of their research, however, I would like to begin with some background.

Over the years Dillman and I have come to suggest that respondents make mistakes responding to visually-administered questionnaires (that is both paper and Web) not only because they do not understand the verbal language of the questionnaire, but because they do not understand the numeric, symbolic, and graphic language as well. (Redline and Dillman, 2002). The verbal language refers to the words, the numeric, the numbers, the symbolic, the symbols, like arrows, and the graphic language is the conduit by which all of the other languages are conveyed and includes the brightness, color, shape, and location of the information. The reason I have reiterated this framework here is because I used it to process the results of the Tourangeau paper, which I think we can all agree are very impressive and very exciting.

I propose rearranging the original content of the Tourangeau et al. paper. I propose discussing the topic Spacing and Position first and Images as Context last.

Spacing and Position or "Location, Location, Location"

Spacing and position is mostly about manipulating the visual element of 'location,' so conceptually it is simpler or more elemental than the remaining topics. Then within Spacing and Position, I propose leading with positional inferences because it tested the heuristic 'left and top' means first, so conceptually it is a good place to begin. Then the remaining two categories, which demonstrate that the physical middle means midpoint, logically follow.

Left and Top Means First

In 1945 Brandt published an eye-movement analysis using a card with squares that were symmetrically located about a locus (Brandt, 1945). The results of his study may be one of the first to demonstrate the heuristic that the left and top of a space is first because he found that all things being equal, subjects' eyes were naturally attracted to the upper left-hand quadrant, and that the least preferred space was the lower right hand quadrant.

However, as we know, items are NOT of equal interest in a questionnaire. There are what we might call conceptually related zones. The work presented today suggests that respondents attribute meaning to the physical space of a conceptual zone. If one overlays physical quadrants on conceptual spaces, one finds that, unless influenced otherwise, respondents tend to process information within a conceptual space in a somewhat predictable order, starting at the top, and left and working across then down, and as shown by Tourangeau and his colleagues, they attribute meaning to the order in which they process this information and its position within this space.

However, surveys begin to get complicated, the moment we deviate from using standard text in a standard size and font because it is clear we really don't understand the effect of the other languages yet—a case in point is the divider line that was used to separate the substantive from the nonsubstantive answer categories. The divider line is an example of using the additional language of symbolic language from the Redline and Dillman framework. Tourangeau and his colleagues discovered that not only did the divider line (this symbolic language) influence respondents' answer choices within the substantive range as predicted; it also had the unintended consequence of attracting respondent's attention to the nonsubstantive response options.

We witnessed similar effects in the Census questionnaire. For instance, the population count question, the most critical question on the questionnaire was supposedly ideally positioned in the upper left-hand corner of the questionnaire, under the heading of Step 1, but the results of an experiment informed by cognitive interviews revealed that respondents were drawn to the large write in space for their name instead, which falls much further down the page under Step 2, Person 1 (Dillman et al., 1996). This is an example in which all of the languages (that is, the verbal, symbolic, numeric, and graphic) conspired to draw respondents' attention away from the most critical question on the questionnaire to a question that was of little importance located further down the page. Thus, the 'left and top means first heuristic' can be overruled in ways that we are beginning to identify and slowly come to understand.

Visibility Principle or "Out of Sight, Out of Mind"

The visibility principle is conceptually related to the last topic in that we are still talking about manipulating the visual element of location, it is just that now we are moving things, relocating them, response options in this case, and definitions in the second, so that it requires additional effort on respondents' part to find them.

I learned as a result of eye-movement research with branching instructions that as little as 9 to 12 characters can place information out of view (Redline and Lankford, 2001), so it is not surprising that if information is placed behind closed doors, so to speak, in drop down boxes, or behind links, respondents will be less likely to see it, but it is great to have more experimental confirmation to this effect.

A recent example I have of the visibility principle comes from my work with the Graduate Student Survey at the National Science Foundation. This survey has both a paper and a Web version—and of the 17 interviews I've conducted with this survey, not one person has ever accessed any of the Web definitions, which are hidden behind a help menu that is itself hidden from view in the upper right hand corner.

And the paper questionnaire is barely any better because the definitions are provided in what can only be called a thicket of text—my point here is that things can be hidden when they appear to be in clear view too.

Images as Context or “A Picture is Worth a Thousand Words”

What we learn in this section of the paper is that pictures can prime memories, in a sense they can provide definitions for respondents, or they can act as standards of comparisons. However, when the picture is in the header, assimilation rather than contrast seems to be the result. This made me think about Feynman’s description of science as a chess game, the fact that is the scientist’s job to figure out the rules of the game from observing it being played. And just when you think you have the rules figured out, someone will castle, and you’ll go, ‘what was that??!!” I began to wonder if a possible explanation was that the picture in the header was less visible and therefore having less of an effect. It was then that I realized that we needed a no picture condition to compare to, just to rule out this possibility. However, I don’t really think that is what this is because I would expect the results to converge in the ‘within header’ condition. I really think this may be a castle--something exciting to look forward to solving.

Impact on Federal Statistics

Before I conclude, I would like to reflect on how I think the research presented today will impact Federal Statistics. On the one hand, I’m excited because I think it is going to go a long way towards bringing attention to these much-deserved issues. And I also think that, as a result, we will be better poised to conduct high quality household Web surveys, especially attitudinal and behavioral surveys. So I most certainly think we should continue in this vein. However, I must also admit to being concerned, concerned that too much emphasis is going towards the design of Web surveys when paper surveys are still the real work horses.

Take for instance the survey I mentioned earlier, the Graduate Student Survey. It is an establishment survey with both a paper and Web component. The Web component gets an 80 percent response rate, so it was taken for granted that respondents were answering the Web survey—but when I went into the field I discovered that 70 percent of the respondents were actually filling out the paper questionnaire first or performing hand calculations on paper, then simply using the Web as a dissemination tool. Thus, paper remains an important tool for respondents, and the true interface between the survey and the respondent is often the paper questionnaire, not the Web.

Also, it is often said that because of the differences between paper and the electronic medium, paper questionnaires cannot simply be transferred to the Web--that the translation process is simply not a one-to-one mapping (e.g., Murphy et al., 2001). But this philosophy ignores the fact that many of these paper questionnaires are so poorly designed to begin with that one would not want to copy them to the Web as they are. A word of caution I have, therefore, is not to overlook the design of paper surveys in our frenzy to design good Web surveys. It may be that the two will need to work in unison, and if we have people working in isolation on one version or the other, which is the direction I see things moving right now, I think we are headed for trouble.

Conclusion

That word of caution aside, I would like to end by saying that without a doubt this is an exceptional set of experiments that provides a great deal of evidence in support of the notion that the visible matters. Although the experiments were carried out in Web surveys, I have tried to demonstrate that there is every reason to believe that the underlying principles hold true for paper questionnaires too, which I would sum up as: location, location, location; out of sight, out of mind; and a picture is worth a thousand words.

References

Brandt, H. (1945), *The Psychology of Seeing,* New York: The Philosophical Library.

Dillman, D., Jenkins, C., Martin, E., and DeMaio, T. 1996. "Cognitive and Motivational Properties of Three Proposed Decennial Census Forms." Report Prepared for the Bureau of the Census. Washington, D.C.

Murphy, E., Nichols, E., Anderson, A., Harley, M., and Pressley, K. 2001. "Building Usability into Electronic Data-Collection Forms for Economic Censuses and Surveys," *Statistical Policy Working Paper 34—Part 4 of 5*. Washington D.C.: The Federal Committee on Statistical Methodology.

Redline, C. and Dillman, D. 2002. "The Influence of Alternative Visual Designs on Respondents' Performance with Branching Instructions in Self-Administered Questionnaires," in Groves, R., Dillman, D., Eltinge, E., and Little, R. (eds.) *Survey Nonresponse*. New York: John Wiley and Sons, Inc.

Redline, C. and Lankford, C. 2001. "Eye-movement Analysis: a New Tool for Evaluating the Design of Visually Administered Instruments (paper and Web)," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Session 4

Robust Small Area Estimation Based on a Survey Weighted MCMC Solution for the General Linear Mixed Model

Hierarchical Bayes Small Area Estimation for Survey Data by EFGL: The Method of Estimating Function-Based Gaussian Likelihood

A. C. Singh, R. E. Folsom, Jr., and A. K. Vaish
RTI International

Abstract

In this paper a new approach representing a generalization of Fay-Herriot (1979) (FH) to unit-level nonlinear mixed models is presented which, like FH, employs data aggregation but through design-weighted estimating functions rather than estimators. Working with estimating functions (EFs) helps to alleviate the problems associated with FH because EFs, in general, can be better approximated by normality even for modest sample sizes, and can always be collapsed, if necessary, to improve the Gaussian approximation and the precision of variance estimates. Also, EFs can be based on unit-level covariate information, and can be specified at the lowest level of aggregation to avoid the problem of internal inconsistency. For hierarchical Bayes (HB) small area estimation, the proposed approach simply replaces the likelihood (typically computed under the assumption of ignorable design) with the estimating function based Gaussian likelihood which does not require ignorability of the design. The method is illustrated by means of a simple example of fitting a HB linear mixed model to data obtained from a nonignorable sample design. Both fixed and random parameters are estimated to construct small area estimates. Different scenarios for nonignorability are considered. MCMC is used for HB parameter estimation.

Key Words: Estimating functions; Pseudo Score Functions; Survey weighted HB; MCMC

1. INTRODUCTION

This research on small area estimation (SAE) was motivated by the problem of fitting generalized linear mixed models to survey data when unit-level covariate information is available. The problem arose in the context of the 1999 National Household Survey on Drug Abuse (NHSDA), see Folsom, Shah, and Vaish (1999). In the NHSDA, one of the outcome variables (y) of interest is past month marijuana use by persons aged 12-17. For this dichotomous variable, one can use as covariates person-level demographic variables, census block-group-level demographic variables, census tract-level demographic and socioeconomic status variables, and inter-censal county-level variables including drug-related arrest, treatment and death rates. For estimating propensity of marijuana use at the state-level (treated as a small area), the following hierarchical Bayes (HB) model similar to the one considered by Folsom et al. may be formulated:

$$\begin{aligned} & y_{ijk} = \mu_{ijk} + \varepsilon_{ijk},\, y_{ijk} \sim \text{Bernoulli}\left(\mu_{ijk}\right) \\ & g\left(\mu_{ijk}\right) = x'_{ijk}\beta + \eta_i + v_{ij} \\ & \eta_i \sim_{iid} N\left(0, \sigma_\eta^2\right), v_{ij} \sim_{iid} N\left(0, \sigma_v^2\right) \\ & \beta \sim U\left(R^p\right), \quad \sigma_\eta^2 \sim IG\left(\nu_0/2, \sigma_{\eta_0}^2/2\right), \sigma_v^2 \sim IG\left(\nu_1/2,\, \sigma_{v0}^2/2\right) \end{aligned} \tag{1.1}$$

where y_{ijk} denotes the observation on the k^{th} individual from the j^{th} cluster (such as a county) in the i^{th} stratum (such as a state), x_{ijk} is the corresponding individual-level covariate vector, and $g(\cdot)$ is the link function (such as the logit). The p-dimensional fixed parameter β has an improper uniform distribution on the p-dimensional space of real vectors, and the variance components σ_η^2, σ_v^2 have nearly flat inverse Gamma priors with very small location and scale parameters $\nu_0 > 0$, $\sigma_{\eta_0}^2 > 0$, $\nu_1 > 0$, $\sigma_{v_0}^2 > 0$. The model errors ε's are independent of each other, and also independent of the random effects η's and v's.

In the context of survey data, the model (1.1) is a super-population model assumed to hold for the finite population U_N of size N. For U_N, let M be the number of strata ($i = 1, \ldots, M$), and N_i be the number of clusters in the i^{th} stratum (j=1, …, N_i), and N_{ij} be the number of individuals in the (i,j)th cluster (k=1, …, N_{ij}). The parameters $\eta_1, \ldots, \eta_M$ are the realized values from $N\left(0, \sigma_\eta^2\right)$. The (random) parameters of interest are the stratum means, μ_i, and the domain means, μ_d where the domain d may cut across strata. Thus,

$$\mu_i = \left(\sum_j \sum_k N_{ijk}\mu_{ijk}\right)\Big/N_i, \quad \mu_d = \sum_i \gamma_{id}\mu_{id}, \tag{1.2}$$

where γ_{id} is the proportion of domain-*d* units in stratum-*i*, and μ_{id} is the mean of the domain-*d* units in stratum-*i*. Other parameters of interest may be the overall mean $\mu\left(= \sum_i \gamma_i \mu_i, \quad \text{where } \gamma_i = \sum_j N_{ij}/N\right)$, and the fixed parameters α, β, σ_η^2 and σ_v^2.

The observed data is a sample (s) of size n from the finite population U_N. If the sample design, $p(s)$, is ignorable for the model (1.1), i.e., the model (1.1) also holds for the sample, s, then the usual HB estimation theory can be applied to s. However, if the design is nonignorable, use of the standard likelihood in the HB framework would lead to a biased posterior distribution, because the model (1.1) cannot be assumed to hold for the sampled data due to selection bias. This is discussed further in Section 2.

In Section 3, we consider existing solutions based on the seminal work of Fay-Herriot's (1979) aggregate-level model, and show how it takes account of the survey design. However, it does have some limitations which are also discussed. Section 4 provides motivation for the alternative proposed solution which is described in Section 5 in the context of a simple example of mixed linear models. The MCMC steps for the proposed HB-SAE method are described in section 6. Sections 7 and 8 describe the simulation experiment and results. The case of mixed nonlinear models is considered in Section 9 which also shows how the proposed method compares with the alternative method of Folsom et al. originally proposed for the NHSDA application. Finally we conclude the paper with some remarks in Section 10.

2. NONIGNORABILITY OF SAMPLE DESIGN

Consider a super-population model which is assumed to hold for the finite population U_N. For the sake of simplicity, we first consider a simple linear mixed model for the observations y_{ij} on the unit j in the ith cluster, ($i = 1,\ldots, M$; j=1, …, N_i). We have

$$y_{ij} = x'_{ij}\beta + \eta_i + \varepsilon_{ij} \tag{2.1}$$

where $\varepsilon_{ij} \sim_{iid} N\left(0, \sigma_\varepsilon^2\right)$, $\eta_i \sim_{iid} N\left(0, \sigma_\eta^2\right)$, β is a p-vector of fixed effects, and x_{ij} is a p-vector of covariates associated with the unit j in the cluster i. Here η_i's are random cluster effects.

We note that in practice it is almost impossible to include in the model all the factor effects (main and interaction) of design covariates such as cluster characteristics that are deemed to be related to the outcome variable y. This happens for several reasons: (i) the need for a parsimonious model, (ii) the need to avoid instability of parameter estimates, (iii) the model should correspond to the analyst's goals, and (iv) some covariates at lower levels are excluded due to unavailability of lower level population totals; these totals are needed in defining finite population parameters.

Since sample selection probabilities may depend on the outcome variables through design covariates, and since all the factor effects due to design covariates may not be controlled in the model, it is difficult to assume that the design can be ignored for the model under consideration. This is why many survey samplers prefer to follow the conventional wisdom of playing it safe by taking the design into account. There are two main scenarios in small area modeling which make the design nonignorable.

Scenario I. Here small areas are, in fact, design strata, and the random effects η_i 's correspond to these strata. Sampling within each stratum is informative in that the sample inclusion probability π_{ij} depends on ε_{ij}. Note that the factors corresponding to design covariates (x_2, say), which are omitted from the model but are correlated with y_{ij}, become naturally part of ε_{ij}. This is easily seen from the following expression for the reduced model $y = E(y \mid x_1) + \varepsilon'$, $\varepsilon' = \left(E(y \mid x_1, x_2) - E(y \mid x_1) + \varepsilon \right)$ when the enlarged model is $y = E(y \mid x_1, x_2) + \varepsilon$.

Scenario II. Here, small areas are like domains, and the random effects η_i 's correspond to these domains. Note that each domain may cut across design strata. In each stratum, sampling may be informative in that the sample inclusion probability of the (i,j)th unit in the hth stratum, $\pi_{h(ij)}$ may depend on η_i or ε_{ij} or both. This is again for the reason that effects of design covariates which are not part of the model covariates x's, become automatically part of the residual, $\eta_i + \varepsilon_{ij}$; here the residual has two components, η_i and ε_{ij}.

Now, in Bayes or hierarchical Bayes estimation, we need specifications of the likelihood, $L\left(y \mid \beta, \eta, \sigma_\varepsilon^2\right)$ and of prior distributions. If L(·) is misspecified, the posterior distribution, $[\theta | y]$, is not correct for parameters of interest θ. (For instance, $\theta_i = A'_{xi}\beta + \eta_i$, is (approximately) the ith area mean where $\sum_j \varepsilon_{ij} / N_i \approx 0$ for large N_i, and A_{xi} is the mean of x for the ith area, i.e., $A_{xi} = T_{xi}/N_i$, $T_{xi} = \sum_j x_{ij}$.) Thus, any characteristic of $[\theta | y]$, in particular the posterior mean, could be (seriously) biased in that

$$E_{\theta|y}\left[\theta - E^*(\theta \mid y)\right] \neq 0 \tag{2.2}$$

where E* denotes the posterior expectation based on the misspecified likelihood.

In the next section, we consider the existing solution of Fay and Herriot (1979, henceforth referred to as FH) in which the sampling design is taken into account by working with the aggregate-level data. Note that for aggregate statistics such as weighted sample totals or means, design-based variances and covariances can be estimated, and their distribution can be approximated as Gaussian. It is difficult in general to specify the distribution of the unit-level data because there is not enough information about the distribution of the N-vector of sample inclusion indicators. In fact, typically, not even all the first order inclusion probabilities are known, let alone second or higher order inclusion probabilities. Some alternative approaches based on modeling of selection probabilities have been proposed by Pfeffermann and Sverchkov (1999). However, with the desirable goal of making minimum modeling assumptions for SAE, a way out might be to do efficient aggregation of data that incorporates unit-level information, and then use sampling weights as in FH, see Section 4. It may be remarked that unlike the census data which is based on nature's selection mechanism of the finite population, the sample from the finite population is based on man's selection mechanism, and hence the sampler knows very well what should not be assumed away. This is probably why the analysis of survey data becomes quite challenging, and thus distinguishes itself from the mainstream of statistics.

3. EXISTING SOLUTION: AGGREGATE LEVEL MODEL OF FH

The work of FH represents a milestone in the history of the development of SAE as it is the first method that takes design into account in small area modeling. The basic idea is to transform the unit-level data (y) to aggregate-level data $(\tilde{y})$ by using the direct small area estimates, $\hat{\theta}_{i,dir}\left(=\sum_{j=1}^{n_i} y_{ij}w_{ij}/w_{i+}\right)$ where w_{ij}s are the (calibrated) design weights, and $w_{i+}\left(=\sum_{j=1}^{n_i} w_{ij}\right)$ is typically equal to N_i due to weight calibration. Thus, in FH, the data is first condensed into M estimates which are modeled as follows. We will consider only Scenario I for the sake of simplicity. For $i=1,\dots,M$; we specify the following

$$\begin{array}{ll} \text{Observation model:} & \hat{\theta}_{i,dir} = \theta_i + e_i,\ \ and \\ \text{Link model:} & \theta_i = A'_{xi}\beta + \eta_i, \end{array} \tag{3.1}$$

where $e \ \dot{\sim}\ N\left(0, diag(V)\right)$, $\ \eta_i \sim_{iid} N\left(0, \sigma_\eta^2\right)$.

Here, V denotes the vector of design-based variance estimates that are regarded as known. In practice, they could be smoothed by suitable modeling; FH used generalized variance functions to smooth V , while Otto and Bell (1995) proposed a parameterization of Cov (e) along with a suitable prior under a Bayesian framework. Even if variance estimates are not smoothed, one could still treat them as known and meet the goal of SAE modeling. The reason for this is that the main goal of SAE modeling is to see whether variances of SAEs after borrowing strength from other areas via modeling can be reduced appreciably in comparison to the variances of direct estimates. Note that under the assumption of unit- level model (2.1), there is another error term involving ε_{ij} in the link model (3.1) given by

$$\begin{aligned}\theta_i &= A'_{xi}\beta + \eta_i + \sum\nolimits_{j=1}^{N_i} \varepsilon_{ij} \Big/ N_i, \\ &\approx A'_{xi}\beta + \eta_i\end{aligned} \tag{3.2}$$

where the term $\sum_j \varepsilon_{ij} / N_i \approx 0$ by SLLN, because N_i is expected to be very large in practice even though n_i may be small. Similarly, the Cov(e) in the observation model involves σ_ε^2 when the covariance is computed under both design and model randomizations, i.e., when the super-population expectation of the design-based covariance is taken. However, it is better to use just the design-based estimate of Cov(e) for several reasons: firstly, the actual computational form for Cov(e) under complex designs may be quite complex involving unknown second order inclusion probabilities, and so computation of its expectation may be prohibitive; secondly, even if the expectation involving σ_ε^2 is computable, one cannot produce good estimates of both σ_ε^2 and σ_η^2 from the aggregate-level data because it is hard to discriminate between them without unit-level data; and thirdly, the design-based estimate of Cov(e) has the desirable property of robustness to departures from the link model.

The Gaussian approximation of $\hat{\theta}_{i,dir} - \theta_i$ in the FH set-up is based on the Central Limit Theorem, and using this, FH proposed empirical Bayes estimators for θ_is. However, if we were interested in HB estimation using the aggregate-level data, the unit-level likelihood L(y|·) can be replaced by the aggregate-level likelihood $\mathrm{L}(\tilde{y}|\cdot)$, and one can then proceed as in Datta and Ghosh (1991).

Although the FH method represents a very important development in SAE methodology for survey data, it does suffer from a few limitations resulting mainly from aggregate-level modeling. Note that when the unit-level model is of interest, there is a loss of efficiency by using an aggregate-level model. This is analogous to the case of using the grouped data mle instead of the raw data mle in chi-square goodness-of-fit tests. While it is true that some loss of efficiency is inevitable when trying to take design into account, the issue under consideration is how to reduce this efficiency loss for unit-level models. Below we list some limitations of the FH approach.

(a) In the aggregate-level modeling approach of FH, unit-level covariate information is not exploited. The more unit-level information is used, the more efficient the resulting estimators are expected to be.

(b) The FH model is specific to the level of aggregation used. If we change the level of aggregation, we get a different model which is not internally consistent with the original model. Note that the exchangeability assumption about η_i 's is specific to the level of aggregation. This inconsistency problem becomes more acute when dealing with nonlinear models either in the mean function of the link model or in the dependent variable of the observation model. For example, with the logit link function, mean at a higher level is not sum of the means at lower levels that make up the higher level of aggregation. In practice, the additive property is clearly desirable. We run into similar problems if $\hat{\theta}_i$ is transformed through a nonlinear function such as $\log \hat{\theta}_i$. Here, an additional problem arises in the definition of $\log \hat{\theta}_i$ when $\hat{\theta}_i = 0$, see e.g. the report on SAIPE models by US Bureau of the Census (1998).

(c) In FH, the Gaussian approximation of $\hat{\theta}_i - \theta_i$ may not be reasonable for small to modest n_i's. This may be more of a concern when dealing with discrete outcome variables.

(d) Finally, smoothed variance estimates V may not be a good approximation for very small n_i 's. Note that, if the direct small area estimates $\hat{\theta}_{i,dir}$ are unstable (this is precisely the reason why we are modeling to borrow strength), then the variance estimates V will, of course, be unstable.

4. MOTIVATION FOR THE ALTERNATIVE SOLUTION

In this paper we propose a generalization of FH to unit-level nonlinear mixed models such that unit-level covariate information is efficiently used as well as some form of data aggregation is used to account for the sample design. Recently, in an innovative attempt to account for the design, Prasad and Rao (1999) derived an aggregate-(or area-) level model for direct estimates from the unit-level model using survey weights, and obtained pseudo-optimal SAEs. It is pseudo in that the design was assumed to be ignorable, and so only the effect of unequal selection probabilities (i.e., sampling weights) was accounted for in the joint design-model variance. Moreover, for estimating variance components, in addition to assuming that the design was ignorable, the unequal weighting effect was also not accounted for. You and Rao (2003) used a similar framework for developing pseudo HB estimates. The above methods, however, are applicable to only linear models because the aggregate-level model for direct estimates is derived from the unit-level model. On the other hand, the method of Folsom et al. (1999) deals with unit-level mixed nonlinear models and develops a HB method using pseudo-likelihood involving survey weights and the corresponding survey weighted estimating functions. However, the method assumes ignorability of the design, and the pseudo likelihood used for HB need not be a valid likelihood; see Section 7 for a brief discussion.

Our goal is to attempt to take full account of the survey design in unit-level modeling, and to develop methods that apply to both linear and nonlinear models. To this end, unlike FH we resort to data aggregation via survey-weighted estimating functions rather than through estimators. Use of survey weighted estimating functions has been implicitly invoked by survey statisticians for a long time in ratio and regression type estimators, see e.g., Fuller (1975), Cassel, Särndal, and Wretman (1976). The pioneering work of Binder (1983) explicitly introduced a general theoretical framework of survey weighted EFs for deriving estimators of super-population parameters, and their asymptotic properties under a given sample design. The optimality of survey-weighted EFs under joint design-model randomization was, however, established by Godambe and Thompson (1986) using the optimality framework of Godambe (1960). As an example, for the simple mixed linear model (2.1), the optimal EFs for β and η_i's have heuristically appealing forms and are given by

$$\begin{aligned} \varphi_{\eta(i)} &= \sum\nolimits_{j=1}^{n_i} \left(y_{ij} - x'_{ij}\beta - \eta_i \right) w_{ij}, \\ \varphi_{\beta} &= \sum\nolimits_{i=1}^{M} \sum\nolimits_{j=1}^{n_i} x_{ij} \left(y_{ij} - x'_{ij}\beta - \eta_i \right) w_{ij} \end{aligned} \tag{4.1}$$

where w_{ij}'s are inverse of the first order selection probabilities π_{ij}'s.

We propose to use the above set of EFs as the starting point for Bayes or HB estimation, i.e., the likelihood would be defined by the distribution of these EFs. Clearly, EFs use unit-level information and they use it efficiently in view of their optimality properties. It is also known that EFs can be better approximated as Gaussian even for modest sample sizes (McCullagh, 1991) because by their very nature, they are simple sums of elementary zero functions, although the elementary functions could be complex by themseleves. Moreover, EFs can be easily collapsed to improve the Gaussian approximation as well as the precision of variance estimates. Notice that the serious problem of internal inconsistency can be avoided by defining the EFs at the lowest level of aggregation. Thus, parameters at higher levels of aggregation can be obtained from the lowest level parameter estimates which serve as building blocks. It should also be noted that, typically in practice, the joint inclusion probabilities $(\pi_{i(jk)})$ of units j and k in stratum i are not available and therefore, survey weighted EFs can't be constructed if they involve cross-product terms, e.g., if they involve double sums within a stratum i. It is, therefore, desirable to specify the model (2.1) so that the error term ε_i 's are i.i.d. which, in turn, gives rise to single sums within strata for survey weighting.

Now, the vector φ of EFs (which involves data and parameters) serves as the condensed input data which after collapsing, if necessary, gives rise to an approximate Gaussian likelihood, L($y*$|β, η, ·) where $y*$ denotes the implicit condensing of information in y via ϕ. Thus, for the unit-level HB analysis, the original likelihood L(y|·) (which would have been based on the ignorable design assumption) is replaced by the estimating function based Gaussian likelihood (EFGL), L($y*$|·) which does not assume ignorability of the design.

5. PROPOSED METHOD (EFGL)

We shall describe the proposed method of estimating function-based Gaussian likelihood (EFGL) in terms of the model (2.1). Suppose, the HB-framework at the census-level is defined as follows:

$$\begin{aligned} &y_{ij} \mid \beta, \eta, \sigma_\varepsilon^2 \sim N\left(x'_{ij}\beta + \eta_i, \sigma_\varepsilon^2\right) \\ &\eta_i \sim_{iid} N\left(0, \sigma_\eta^2\right), \quad \beta \sim U\left(R^p\right) \\ &\sigma_\eta^2 \sim IG\left(\nu_0/2, \sigma_{\eta_0}^2 \mid 2\right), \quad \sigma_\varepsilon^2 \sim U\left(0, \infty\right). \end{aligned} \tag{5.1}$$

Here an attempt is made to specify the priors to make them as noninformative as possible, and thus making the HB framework as objective as possible. Thus, the p-vector β of regression coefficients is assumed to have an improper uniform prior on the p-dimensional Euclidean space. However, this does not affect the propreitory of the posterior of β . For variance component σ_η^2, choice of the inverse Gamma as prior is computationally convenient because of its conjugate nature, and we can choose the shape parameter $\left(\nu_0/2\right)$ and the scale parameter $\left(\sigma_{\eta_0}^2/2\right)$ as very small positive numbers to make it nearly noninformative. The prior for σ_ε^2, however, is improper like that of the mean parameter β, because in EFGL, as will be seen later, we introduce a separate EF, $\varphi_{\sigma^2(\varepsilon)}$, for σ_ε^2 which treats σ_ε^2 as a mean parameter. It turns out as expected and as in the case of FH that σ_ε^2 is not functionally part of the V-C matrix Σ_ϕ of φ when a suitable design-based estimate of Σ_φ is substituted. So we need to add an extra EF if the estimation of σ_ε^2 is also of interest. It may be noted that there is quite a bit of flexibility in the EF framework in that all the pieces of information deemed important can be incorporated by augmenting the vector φ.

Now, the EFGL method will be defined for Scenario I in which small areas are strata. The EFs $\varphi_{\eta(i)}$ and φ_β were defined earlier by (4.1). Further suppose,

$$\varphi_{\eta(i)} \sim_{approx} N\left(0, V_{\eta(i)}\right), \quad \varphi_{\beta} \sim_{approx} N\left(0, V_{\beta}\right) \text{ and } Cov\left(\varphi_{\beta}, \varphi_{\eta(i)}\right) = C_{\beta\eta(i)}. \tag{5.2}$$

Next define

$$\tilde{\varphi}_{\beta} = \varphi_{\beta} - \Sigma_i a_i \varphi_{\eta(i)}, \quad a_i = C_{\beta\eta(i)} \big/ v_{\eta(i)}$$

which implies that $\tilde{\varphi}_{\beta}$ is uncorrelated with $\varphi_{\eta(i)}$'s. It should be remarked that if the model (2.1) has an intercept β_0, then $\varphi_{\beta 0} = \sum_i \varphi_{\eta(i)}$ implying that $\tilde{\varphi}_{\beta 0} = 0$. We, therefore, drop one element from φ_{β} corresponding to β_0. However, we shall continue to use φ_{β} to denote the reduced vector of dimension $p-1$. Further, since

$$Cov\left(\tilde{\varphi}_{\beta}\right) \equiv \tilde{V}_{\beta} = V_{\beta} - C_{\beta\eta} V_{\eta}^{-1} C'_{\beta\eta}, \quad V_{\eta} = diag\left(V_{\eta(1)}, ..V_{\eta(M)}\right). \tag{5.3}$$

We have

$$\tilde{\varphi} = \left(\varphi_{\eta}, \tilde{\varphi}_{\beta}\right)' \sim_{approx} N_{M+p-1}\left(0, \tilde{V}_{\varphi}\right), \quad \tilde{V}_{\varphi} = \text{ blockdiag } \left(V_{\eta}, \tilde{V}_{\beta}\right) \tag{5.4}$$

and the EFG log-likelihood is given by

$$\ell\left(y^* \mid \beta, \eta\right) = \text{ const} - \frac{1}{2}\left(\frac{\varphi_{\eta(1)}^2}{v_{\eta(1)}} + \ldots + \frac{\varphi_{\eta(M)}^2}{v_{\eta(M)}} + \tilde{\varphi}'_{\beta} \tilde{V}_{\beta}^{-1} \tilde{\varphi}_{\beta}\right). \tag{5.5}$$

In the above EFGL, the covariance matrix $\tilde{V}_{\varphi}$ is design-based. This matrix may, in general, depend on unknown parameters which can be evaluated at their current values in the MCMC samples. It may be noted that there is, in fact, a second component involving σ_{ε}^2 when the V-C matrix of $\tilde{\varphi}$ is computed under joint design-model randomization. However, it is negligible in comparison to the first term, $\tilde{V}_{\varphi}$, under the usual assumption of n_i<<N_i. It should also be emphasized that, in practice, some collapsing of $\varphi_{\eta(i)}$'s may often be required because the corresponding n_i's (which are random under Scenario II) may be small. We may need this collapsing to improve the Gaussian approximation, as well as to improve the precision of the estimate $\tilde{V}_{\varphi}$. The effect of EF-collapsing on η_i-estimates is that all the prior estimates of θ_i's $\left(\theta_i = A'_{xi}\beta + \eta_i\right)$ that are part of a given collapsed EF, are shrunk toward the direct estimate of the corresponding collapsed small area. It is, therefore, important to choose EF-collapsing partners carefully so that they have similar η_i's both in magnitude and sign. To this end, one can make a decision based on substantive considerations. However, in practice, as a yardstick one can use $\hat{\eta}_{i,HB}^{(0)}$ obtained under the ignorability assumption. Once it is decided which η_i's would be used in EF-collapsing, one can construct a new census EF under the assumption of common η_i's for this set, and then employ survey weighting to get the appropriate collapsed EF.

If estimation of σ_ε^2 is also of interest, we add an extra EF as mentioned earlier. It is again motivated by census EF, and is given by

$$\varphi_{\sigma^2(\varepsilon)} = \sum_i \sum_j \left(\left(y_{ij} - x_{ij}'\beta - \eta_i \right)^2 - \sigma_\varepsilon^2 \right) w_{ij} \sim_{approx} N\left(0, V_{\sigma^2(\varepsilon)}\right). \tag{5.6}$$

Note that in FH, although σ_ε^2 is not made explicitly part of the model, it could be done so by taking expectation of the design-based variance V. However, as mentioned earlier, using aggregate-level data $\hat{\theta}_{i,dir}$, it would be difficult to discriminate very well between the two variance components σ_η^2 and σ_ε^2.

With the specification of EFGL, estimation of parameters $\left(\eta, \beta, \sigma_\eta^2, \sigma_\varepsilon^2\right)$ can proceed in the HB setup using MCMC steps. The next section gives details of full conditional posterior distributions needed for MCMC. Although so far, we have considered only Scenario I, the case of Scenario II is somewhat analogous. The main difference is that the V-C matrix of φ_η is no longer diagonal, and so the form of the EFGL is not as simple. However, full conditional posterior distributions (Section 6), can be derived easily by first orthogonalizing φ_β with respect to φ_η, and then for each i, orthogonalizing $\varphi_{\eta(i)}$ with all other $\varphi_{\eta(i')}$, $i' \neq i$.

6. MCMC FOR THE PROPOSED HB-SAE

For the Scenario I, the MCMC steps for finding full conditionals can be defined as follows. It is assumed that the regularity conditions for the convergence of the MCMC steps toward a stationary distribution hold.

<u>Step I</u>. $[\beta \mid y^*, \eta]$

We note that under the vague uniform prior for β, the posterior of β is simply proportional to the likelihood, and is given by

$$\log[\beta \mid \cdot] = const - \frac{1}{2}\left[\sum_{i=1}^{M} \varphi_{\eta(i)}^2 \big/ v_{\eta(i)} + \tilde{\varphi}_\beta' \tilde{V}_\beta^{-1} \tilde{\varphi}_\beta \right]. \tag{6.1}$$

Since the kernel of the log-likelihood involves first and second powers of β, one can complete after some algebra the quadratic form in β. This implies that [β|·] is exact Gaussian with mean and V-C matrix given respectively by the mode and curvature (at mode of the above kernel function if it depends on β). Thus,

$$\left[\beta \mid y^*, \eta\right] = N_{p-1}\left[\hat{\beta}_{\text{mode}},\ \Sigma_{\psi(\beta)}^{-1}\right], \tag{6.2}$$

where $\hat{\beta}_{\text{mode}}$ solves the estimating equation $\psi_\beta = 0$,

$$\psi_\beta = \left(\partial/\partial\beta\right)\log L\left(y^*\middle|\beta,\eta\right) = \left(\sum_{i=1}^{M}\varphi_{\eta(i)}x_{i+w}\right)\Big/V_{\eta(i)} + \left(\tilde{X}'W\tilde{X}\right)\tilde{V}_\beta^{-1}\tilde{\varphi}_\beta. \tag{6.3}$$

where $\tilde{X}'W\tilde{X} \equiv X'WX - \sum_i a_i x'_{i+w}$, $x_{i+w} = \sum_i x_{ij}w_{ij}$, and $X'WX = \sum\sum x_{ij}x'_{ij}w_{ij}$. It is seen that similar to generalized least squares, $\hat{\beta}_{\text{mod}e}$ can be obtained in a closed form. The V-C of ψ_β is easily obtained as

$$\Sigma_{\psi(\beta)} = -E\left[\partial\psi_\beta\Big/\partial\beta\right] = \sum_{i=1}^{M} x_{i+w}x'_{i+w}\Big/V_{\eta(i)} + (\tilde{X}'W\tilde{X})\tilde{V}_\beta^{-1}(\tilde{X}'W\tilde{X}) \tag{6.4}$$

<u>Step II</u>. $\left[\eta_i \mid \eta_{i'}, \beta, y^*, \sigma_\eta^2, i' \neq i\right], \quad i = 1, \dots, M$

Since the posterior [η_i|·] is proportional to the product of the likelihood and the prior, we have

$$\log\left[\eta_i \mid \cdot\right] = const. - \frac{1}{2}\left[\sum_{i=1}^{M}\varphi_{\eta(i)}^2\Big/v_{\eta(i)} + \tilde{\varphi}'_\beta\tilde{V}_\beta^{-1}\tilde{\varphi}_\beta + \eta_i^2\Big/\sigma_\eta^2\right]. \tag{6.5}$$

As in Step I, the kernel on the right hand side of (6.5) involves first and second powers of η_i, and one can complete the square in η_i. Therefore, $[\eta_i|\cdot]$ is also exact Gaussian with mean and variance given by the mode and curvature. That is,

$$\left[\eta_{\text{i}} \mid \cdot\right] = N\left[\hat{\eta}_{i,\text{mode}}, \sigma_{\psi(\eta(i))}^{-2}\right] \tag{6.6}$$

where

$$\begin{gathered}\hat{\eta}_{i,\text{mode}} \text{ solves } \psi_{\eta(i)} = 0,\\ \psi_{\eta(i)} = \left(\partial/\partial_{\eta(i)}\right)\log\left[\eta_i \mid \cdot\right] = \varphi_{\eta(i)}w_{i+}\big/v_{\eta(i)} + \tilde{x}'_{i+w}\tilde{V}_\beta^{-1}\tilde{\varphi}_\beta - \eta_i\big/\sigma_\eta^2.\end{gathered} \tag{6.7}$$

where $\tilde{x}_{i+w} = x_{i+w} - a_i w_{i+}$. Again as in the case of β, $\hat{\eta}_{i,\text{mode}}$ has a closed form. The variance $\sigma_{\psi(\eta(i))}^2$ is obtained as

$$\sigma_{\psi(\eta(i))}^2 = -E\left(\left(\partial/\partial_{\eta(i)}\right)\psi_{\eta(i)}\right) = w_{i+}^2\big/V_{\eta(i)} + \tilde{x}'_{i+w}\tilde{V}_\beta^{-1}\tilde{x}_{i+w} + \sigma_\eta^{-2} \tag{6.8}$$

It is interesting to note that $\psi_{\eta(i)}$ and $\sigma_{\psi(\eta(i))}^2$ coincide with the usual BLUP theory when the design is ignorable and w_{ij} = w (a constant). To see this, note that under the ignorality assumption,

$$a_i = C_{\beta\eta(i)} \Big/ V_{\eta(i)} = Cov\left(\sum_i \sum_j x_{ij} e_{ij} w_{ij}, \sum_j e_{ij} w_{ij} \Big/ \sigma_\varepsilon^2 \sum_j w_{ij}^2\right) = \sigma_\varepsilon^2 \sum_j x_{ij} w_{ij}^2 \Big/ \sigma_\varepsilon^2 \sum_j w_{ij}^2 \,, \tag{6.9}$$

where $e_{ij} = y_{ij} - x_{ij}'\beta - \eta_i$. Thus, assuming $w_{ij} = w$, we have $a_i = \sum_j x_{ij} / n_i$ and $x_{i+w} - a_i w_{i+} = 0$. Also,

$$w_{i+}^2 \Big/ v_{\eta(i)} = w_{i+}^2 \Big/ \sigma_\varepsilon^2 \sum_j w_{ij}^2 = n_i \Big/ \sigma_\varepsilon^2 \tag{6.10}$$

The reduced forms of $\psi_{\eta(i)}$ and $\sigma_{\psi(\eta(i))}^2$ are $\psi_{\eta(i)} = \sum_j e_{ij} / \sigma_\varepsilon^2 - \eta_i / \sigma_\eta^2$, and

$$\sigma_{\psi(\eta(i))}^2 = n_i \big/ \sigma_\varepsilon^2 + 1 \big/ \sigma_\eta^2 = \frac{\sigma_\eta^2 + \sigma_\varepsilon^2 / n_i}{\sigma_\eta^2\, \sigma_\varepsilon^2 / n_i} \tag{6.11}$$

which implies that

$$\hat{\eta}_{i,\text{BLUP}} = \left(\frac{n_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\eta^2}\right)^{-1} \frac{\sum_j \left(y_{ij} - x_{ij}'\beta\right)}{\sigma_\varepsilon^2} = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2 / n_i} \sum_j \left(y_{ij} - x_y'\beta\right) \Big/ n_i \tag{6.12a}$$

and

$$E\left(\hat{\eta}_{i,\text{EFGL}} - \eta_i\right)^2 = \sigma_{\psi(\eta(i))}^{-2} = \frac{\sigma_\eta^2\, \sigma_\varepsilon^2 / n_i}{\sigma_\eta^2 + \sigma_\varepsilon^2 / n_i} = E\left(\hat{\eta}_{i,\text{BLUP}} - \eta_i\right)^2 \tag{6.12b}$$

<u>Step III</u>. $\left[\sigma_\eta^2 \mid \eta\right]$

In view of the conjugate nature of the prior, the conditional posterior also has the inverse Gamma distribution, and is given by

$$\left[\sigma_\eta^2 \mid \eta\right] = IG\left[\left(\nu_0 + M\right) \Big/ 2, \left(\sigma_{\eta_0}^2 + \sum_i^M \eta_i^2\right) \Big/ 2\right] \tag{6.13}$$

which implies that the conditional posterior mean of σ_η^2,

$$E\left[\sigma_\eta^2 \mid \eta\right] = \left(M\left(\Sigma \eta_i^2 / M\right) + \left(\nu_0 - 2\right)\sigma_{\eta_0}^2\right) \Big/ \left(M + \nu_0 - 2\right) \tag{6.14}$$

It follows that the unconditional posterior mean of σ_η^2, i.e. $E\left[\sigma_\eta^2 \mid \tilde{y}\right]$ is obtained by the average of MCMC realizations after convergence. This posterior mean is known to be approximately equal to the REML estimator for large M, see Kass and Steffey (1986), and Singh, Stukel, and Pfeffermann (1996).

Next, if σ_ε^2 also needs to be estimated, then logL gets modified due to inclusion of the EF $\varphi_{\sigma^2(\varepsilon)}$. It is easily seen that in Step III, $[\eta_i\,|\,\cdot]$ can be obtained using the Metropolis-Hastings (MH)-step with a proposal distribution given by the earlier closed form of $[\eta_i\,|\,\cdot]$ where σ_ε^2 is not part of the likelihood.

Now, for estimating σ_ε^2, we add a fourth step.

Step IV. $\left[\sigma_\varepsilon^2\,|\,\cdot\right]$

It is similar to $[\beta\,|\,\cdot]$ because σ_ε^2 is treated like a mean parameter via EF. So,

$$\left[\sigma_\varepsilon^2\,|\,\cdot\right] = Const \times N\left(\hat{\sigma}^2_{\varepsilon,\,\mathrm{mode}}, V_{\sigma^2(\varepsilon)}\big/ w_{++}^2\right) I_{\left\{\sigma_\varepsilon^2 > 0\right\}} \tag{6.15}$$

where $\hat{\sigma}^2_{\varepsilon,\,\mathrm{mode}} = \sum_i^M \sum_j^{n_i} \left(y_{ij} - x'_{ij}\beta - \eta_i\right)^2 w_{ij}\big/ w_{++}$, and $w_{++} = \sum_i \sum_j w_{ij}$ is typically constant in practice due to weight calibration.

Before moving to the next section, we remark that in the HB framework, to get a reasonable shrinkage of the prior estimates of η_i toward the direct estimates, we need most of the η_i's manifested in the sample. If the sampling design is such that this is not the case (e.g., if η_i's correspond to random PSU effect), then we are faced with an undesirable scenario in which there is hardly any shrinkage of prior estimates of η_i's. It is interesting to note an analogy of the above situation with the model-based estimation in survey sampling under the prediction approach, where the model-based predictor of the unobserved part of the population is simply given by the synthetic estimator.

7. SIMULATION EXPERIMENT

We design our study along the lines of Pfeffermann et al. (1998). Consider a universe of $i = 1, \cdots, M$ strata (small areas) where $M = 100$ and let N_i denote the number of population members in stratum-i. In this simulation experiment, we set $N_i = N_0\,(1 + \exp(u_i^*))$ where N_0 is a constant and u_i^* is obtained by truncating $u_i \sim N(0, 0.2)$ at $\pm\sqrt{0.2}$. For simplicity, we consider a single covariate super-population linear mixed model $y_{ij} = \beta_0 + x_{ij}\ \beta_1 + \eta_i + \varepsilon_{ij}$ where $\beta_0 = 0.5$, $\beta_1 = 1$, $\eta_i \sim N(0, 0.2)$, $\varepsilon_{ij} \sim N(0, 4)$, and $j = 1, \cdots, N_i$. The covariate $x_{ij} = \upsilon_i + \delta_{ij}$ where $\upsilon_i \sim N(0, 0.1)$ and $\delta_{ij} \sim N(0, 1)$. We generate $K = 150$ population level data sets with common x_{ij} and N_i where N_i's are generated using N_0 =3000. Note that the substratum sizes vary over the 150 populations. We selected two samples from each of these populations. The first sample was selected in such a way that the design was ignorable. The second sample was selected so that the design was nonignorable.

To select a sample with an ignorable design, we further stratify the stratum-i population into two substrata Ω_{i+} with $x_{ij} > 0$ and Ω_{i-} with $x_{ij} \le 0$. To select a sample with nonignorable design, we stratify the stratum-i population into two substrata Ω_{i+} with $\varepsilon_{ij} > 0$ and Ω_{i-} with $\varepsilon_{ij} \le 0$. Let N_{i+}, N_{i-} denote the sizes of these substrata and n_{i+}, n_{i-} denote the sizes of the simple random samples selected without replacement from these strata, respectively. Note that the substratum sizes vary across populations. Let $N = \sum_{i=1}^{100} N_i$ and $n = \sum_{i=1}^{100} n_i$ where $n_i = n_{i-} + n_{i+}$. For 150 populations, we generate the corresponding 150 samples. In our simulation experiment, $N = 628897$, $n_{i-} = 60$ and $n_{i+} = 20$ so that we have a sample of size 80 for each small area with a total sample of size 8000.

In our simulation study, we compare EFGL, FH, unweighted HB, and PHB (Pseudo-hierarachical Bayes method of You and Rao, 2003) solutions by comparing average posterior means and standard deviations of the parameters of interest. We also compare average 95% prediction interval coverage probabilities as well as the average lengths of 95% prediction intervals. These averages are taken over 150 replications corresponding to populations with varying η_i's. The comparisons are made for samples generated under ignorable and nonignorable designs. For the FH method, we used a HB-version obtained from EFGL by transforming the unit-level auxiliary information to the aggregate-level, i.e., replacing x_{ij} with $\bar{X}_i = (\sum_{j=1}^{N_i} x_{ij}) \div N_i$. For the PHB method, we used version 2 of You and Rao (2003).

For each sample ($s = 1,\cdots,150$), using Gibbs sampling technique, we generate 10,000 MCMC samples for each of the model parameters, namely $\beta_0, \beta_1, \eta_1, \ldots \eta_{\mathrm{M}}$, and σ_η^2. These MCMC samples are tested for convergence criterion using CODA (Convergence Output Data Analysis software). First 1000 MCMC samples are deleted for "burn-in" period and from the rest of the 9000 MCMC samples we selected every ninth sample to minimize any auto-correlation among samples, yielding a final MCMC sample of size 1000.

Let $\theta_{sc} = (\beta_{0sc}, \beta_{1sc}, \eta_{isc}, \sigma_{\eta sc}^2)$ denote the parameter values from the c-th MCMC cycle corresponding to the s-

th sample. In Table 1, the average posterior mean of θ is defined as $(\sum_{s=1}^{150}\sum_{c=1}^{1000}\theta_{sc})\div(1000\times150)$ and the average posterior standard deviation of each element of θ_{sc} is defined as the square root of $(\sum_{s=1}^{150}\sum_{c=1}^{1000}(\theta_{sc}-\overline{\theta}_s)^2)\div(1000\times150)$ where $\overline{\theta}_s=(\sum_{c=1}^{1000}\theta_{sc})\div1000$. Let $\Theta_{isc}=\beta_{0sc}+\overline{X}_i\ \beta_{1sc}+\eta_{isc}$ denotes the small area estimate from the s-th sample for the i-th area using the c-th MCMC cycle where $\overline{X}_i=(\sum_{j=1}^{N_i}x_{ij})\div N_i$. Also, define $\Theta_{is}^*=\beta_0+\overline{X}_i\ \beta_1+\eta_{is}$ where η_{is} is the true value of η_i for the s-th population. Let L_{is} and U_{is} denote 2.5 and 97.5 percentiles of the posterior distribution of Θ_{is} obtained from 1000 MCMC samples of Θ_{isc}.

Define $\psi_{is}=\begin{cases}1 & \text{if} \quad \Theta_{is}^*\in[L_{is},U_{is}] \\ 0 & \textit{otherwise.}\end{cases}$

The coverage probability distribution characteristics given in Tables 2 are obtained from the distribution of 100 area-i specific values of $(\sum_{s=1}^{150}\psi_{is})\div150$.

8. SIMULATION RESULTS

Tables 1 and 2 summarize the simulation results obtained from the ignorable sample design, whereas Tables 3 and 4 present the corresponding results for the nonignorable samples. In Table 1, average posterior means and standard deviations for the EFGL method are compared with solutions from a HB version of the FH model, PHB and unweighted solutions for the ignorable sample design. Since the model holds in the sample, the unweighted solution is expected to be the most efficient solution. The average posterior means for all four methods are very close to each other. The average posterior standard deviation of β_1 for the FH model is approximately 13 times larger than the other methods. This is due to the fact that the FH solution uses aggregate-level covariate information. However, the average posterior standard deviations of β_0 and σ_η^2 for all the solutions are very close to each other.

In Table 2, 95% prediction interval coverage probabilities for the EFGL solution are compared with the FH, PHB, and unweighted HB solutions coverage probabilities. The coverage probabilities for all solutions are very close. However, the prediction intervals for the FH solution are 16% wider than the EFGL solution, which is expected, since the EFGL solution utilizes unit-level covariate information whereas the FH solution uses aggregate-level covariate information. The unweighted HB method, being the most efficient for the ignorable sample design, results in prediction intervals that are approximately 10% shorter than the EFGL solution.

For the nonignorable sample design, Table 3, shows that the average posterior mean for β_0 from the unweighted solution is heavily biased (0.1043 vs 0.5) due to the fact that we over- sample the Ω_{i-} substrata. On the other hand, the average posterior means for the FH, EFGL and PHB solutions are very close to each other. The average posterior standard deviations of β_0 and σ_η^2 for all four solutions are also close to each other whereas the average posterior standard deviation for β_1 from the EFGL, PHB and unweighted solutions are more than 12 times smaller than the solution from the FH model.

From Table 4 (for the nonignorable sample design), we see that 95% coverage probabilities for the EFGL solution and FH solution are very close to each other whereas the coverage probabilities for the PHB solution are approaching 1 and the coverage probabilities for unweighted solution are close to 0. The unweighted method performed very poorly due to the heavily biased estimate of β_0. It suggests that for our nonignorable samples, the PHB solution substantially overestimates the SAE posterior variances. The prediction intervals for the FH, PHB, and unweighted solutions are respectively 86%, 52%, and 32% wider than the EFGL solution. The inefficiency in the FH solution is expected for the reasons mentioned earlier, since the EFGL solution utilizes unit-level covariate information whereas the FH solution uses aggregate-level covariate information.

9. MIXED NONLINEAR MODELS: LOGISTIC CASE

The method of EFGL introduced in Section 5 for finding HB-SAE in the context of mixed linear models can be easily applied to mixed nonlinear models, the only difference being that full conditional posteriors of β and η have no longer analytic solutions. Therefore, as expected, the method gets more computer intensive. To illustrate the ideas, we consider a simpler version of the mixed logistic model (1.1) given by:

$$\begin{aligned}
&y_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad y_{ij} \sim \text{Bernoulli} \\
&\log it\left(\mu_{ij}\right) = x'_{ij}\beta + \eta_i \qquad\qquad (7.1) \\
&\eta_i \sim_{iid} N\left(0, \sigma_\eta^2\right), \quad \beta \sim U\left(R^p\right), \quad \sigma_\eta^2 \sim IG\left(v_0/2, \sigma_{\eta_0}^2/2\right).
\end{aligned}$$

The EFs in this case remain similar to the linear case except that the elementary zero functions (or the residuals) $y_{ij} - \mu_{ij}$, are complex due to the nonlinear form of μ_{ij}'s. Observe that EFs continue to be simple linear functions of elementary zero functions, and hence they behave well in terms of Gaussian approximations. The EFs for the logistic case under Scenario I are given by

$$\begin{aligned} \varphi_{\eta(i)} &= \sum_{j=1}^{n_i} \left(y_{ij} - \mu_{ij}\right) w_{ij} \sim_{approx} N\left(0, V_{\eta(i)}\right) \\ \varphi_{\beta} &= \sum_{i} \sum_{j} x_{ij} \left(y_{ij} - \mu_{ij}\right) w_{ij} \sim_{approx} N\left(0, V_{\beta}\right) \end{aligned} \tag{7.2}$$

We can orthogonalize φ_{β} with respect to $\varphi_{\eta(i)}$'s as before. Also with the intercept model, φ_{β} corresponding to the intercept should be dropped because of its linear dependence on $\varphi_{\eta(i)}$'s. Now, the likelihood, $L\left(y^* \mid \beta, \eta\right)$ can be approximately specified as before, but the MCMC steps are modified as follows:

<u>Step I</u>. $\left[\beta \mid y^*, \eta\right]$

Since the sample is typically very large, the full conditional posterior can be well approximated by

$$\left[\beta \mid y^*, \eta\right] \sim N\left(\hat{\beta}_{\text{mode}}, \Sigma_{\psi(\beta)}^{-1}\right) \tag{7.3}$$

where $\hat{\beta}_{\text{mode}}$ solves $\psi_{\beta} = 0,\quad \Sigma_{\psi(\beta)} = -E\left(\left(\partial/\partial\beta\right)\psi_{\beta}\right)$,

$$\begin{aligned} \psi_{\beta} &= \left(\partial/\partial\beta\right) \log L\left(y^* \mid \beta, \eta\right) \\ &= \sum_{i=1}^{M} \varphi_{\eta(i)} \sum_{j=1}^{n_i} x_{ij} \mu_{ij} \left(1 - \mu_{ij}\right) w_{ij} \Big/ V_{\eta(i)} - \left(\sum_{i} \sum_{j} x_{ij} x_{ij}' \mu_{ij} \left(1 - \mu_{ij}\right) w_{ij} - \sum_{i} a_i \mu_i (1 - \mu_i) x_{ij} w_{ij}\right) \tilde{V}_{\beta}^{-1} \tilde{\varphi}_{\beta} \end{aligned} \tag{7.4}$$

Note that unlike the linear case, $\hat{\beta}_{\text{mode}}$ does not have an analytic form. Also note that instead of the approximate posterior (7.3), one can get realizations from an exact posterior by using the MH step within MCMC in which (7.3) can be used as a proposal.

<u>Step II</u>. $\left[\eta_i \mid \eta_{i'}, \beta, y^*, \sigma_{\eta}^2\right],\quad i = 1, \ldots, M$.

As mentioned earlier, this again does not have an analytic solution. We could use MH with mle/prior for the proposed distribution. In other words, solve $\psi_{\eta(i)} - \sigma_{\eta}^{-2} \eta_i = 0$ to get $\hat{\eta}_{i,\text{mle-adj}}$, where $\psi_{\eta(i)} = \left(\partial/\partial\eta_i\right) \log L\left(y^* \mid \beta, \eta\right)$, and use $N\left(\hat{\eta}_{i,\text{mle-adj}}, \left(\sigma_{\psi(\eta(i))}^2 + \sigma_{\eta}^{-2}\right)^{-1}\right)$ as the proposal distribution where

$\sigma^2_{\psi(\eta(i))} = -E\left[\partial\psi_{\eta(i)}/\partial\eta_i\right]$.

Step III. $\left[\sigma^2_\eta \mid \eta\right]$

We get the same result as in the linear case. Note that Step IV for $\left[\sigma^2_\varepsilon \mid \cdot\right]$ is not needed because σ^2_ε is a known function of μ_{ij} in the logistic case.

We now consider the work of Folsom et al. (1999) mentioned earlier in Sections 1 and 2 which is related to the proposed EFGL method. For the logistic model, they constructed a pseudo log-likelihood (from the Bernoulli likelihood at the census-level) involving design weights. For this purpose, survey weights were scaled such that they sum to the effective sample size obtained by using the design effect within each area i. The design effect was, however, based only on the effect of unequal weighting under the working assumption of ignorability of the design. In other words, effects of stratification, clustering, and multistage designs were ignored.

Under their pseudo-likelihood approach, the score function for η_i involves $\varphi_{\eta(i)}$ multiplied by a scale adjustment for weights. This pseudo score function in conjunction with the prior information gives the appropriate prior-adjusted pseudo-mle for random effects. This prior-adjusted pseodo-mle along with its variance can be used for defining a Gaussian proposal distribution for the MH step in finding the full conditional posterior of η_i. In the case of β, the actual pseudo score function obtained from the pseudo likelihood was, however, not used, but a somewhat modified pseudo score function, namely ϕ_β obtained from the census likelihood was used as it has the appealing property of self-calibration or benchmarking explained later on. Note that the actual pseudo score function for β is not proportional to ϕ_β because of weight scaling. However, $\hat{\beta}_{\text{pseudo mle}}$ obtained by solving $\varphi_\beta = 0$, and the associated sandwich V-C matrix $\left(\partial\varphi_\beta/\partial\beta'\right)^{-1}\Sigma_\phi\left(\partial\varphi_\beta'/\partial\beta\right)^{-1}$ used respectively as the mode and curvature of a Gaussian distribution is likely to be close to the conditional posterior based on the actual pseudo-score function for β. Here the V-C matrix Σ_φ is computed under the working assumption of ignorable designs, and thus reflects only unequal weighting effect. It may be noted that use of the sandwich V-C (and not the pseudo information) matrix is appropriate because the likelihood is pseudo.

For computing, $[\sigma_\eta^2 \mid \cdot]$, the distribution of Step III of EFGL was used. Thus, the above pseudo-likelihood approach has some similarity with the proposed EFGL. The main differences are that the likelihood is pseudo which need not be valid, and the working assumption of ignorable sample design may not be reasonable. In EFGL, the likelihood is based on EFs and approximated by a valid Gaussian likelihood where the covariance matrix takes full account of the design. However, in the NHSDA application, it was observed that the MCMC method for the pseudo-likelihood approach did converge and provided good results. Also, it can be shown that HB-SAE estimates based on the pseudo-likelihood approach have the desirable property of approximate self-calibration or benchmarking because SAEs obtained directly from pseudo score functions are very similar for large samples to the direct SAEs which are,of course, design-consistent. Thus, SAEs for big states will be approximately equivalent to the direct estimates. Also, aggregates of SAE estimates are nearly calibrated to the national direct estimates. By contrast, estimates resulting from the method of EFGL, although design-consistent, need to be modified to achieve benchmarking to direct estimates for areas with very large samples, see e.g., Singh and Folsom (2001).

10. CONCLUDING REMARKS

The method of EFGL was developed to exploit unit-level information, to take full account of the survey design, and to have a valid (approximate) likelihood for the HB-SAE methodology for generalized linear mixed models. It generalizes the aggregate-level model of FH (1979), and the pseudo-likelihood approach of Folsom et al. (1999). There are essentially two main ideas in EFGL, namely, the data aggregation via EFs and EF-collapsing. The main reason for EF-collapsing is to improve Gaussian approximation, and the secondary purpose is to improve the variance estimate's precision. In practice, it may be preferable to use separate modeling to make variance estimates more stable. However, even if variance estimates are not precise, it is often of interest, in practice, to see how much variance reduction can be realized through SAE modeling.

The idea of data aggregation in EFGL is somewhat similar to that of FH except it tries to take advantage of the unit-level information as much as possible. Since EFGL uses more information than FH, the resulting estimates are expected to be more efficient than those from FH. In particular, for the case of simple linear mixed models (2.1) with known variance components, it can be easily shown analytically that precision of the estimates of fixed effects (β) can be improved substantially in the case of unit-level models if the covariates (x_{ij}) have sufficient variability within areas. There is also some gain in efficiency of random effect

(η_i 's) estimates. However, if η_i 's are also defined as coefficients of suitable covariates (z_{ij} 's) as in the case of random regression coefficients, then high efficiency gains in estimating random effects can also be realized if there is sufficient variability in z_{ij} 's within areas.

We remark that the problem of HB-SAE arose in the context of NHSDA-SAE application where it was desired to fit a mixed logistic linear model. This was a daunting SAE application task with a very large data set and many covariates which was addressed by Folsom et al. (1999). Note that it was not possible to use any existing software for this task .

The ideas underlying the proposed method of EFGL are quite general, and the method is applicable to general nonlinear mixed models for survey data. However, it does have some limitations which the user should keep in mind: (i) Some loss of efficiency is inevitable due to data aggregation, and EF-collapsing. This is the price we pay for not having enough information about the likelihood of the sampled data, and by not being able to ignore the sample design. (ii) The EF-collapsing may be needed for the Gaussian approximation. In practice, it is better to avoid it if possible as it doesn't distinguish much between the areas involved in collapsing. At the design stage, one can take measures to ensure a sufficient number of observation in each small area in order to avoid EF-collapsing. It may be noted that one only needs a modest size of the realized sample in small areas for Gaussian approximation of EFs. However, SAEs are still needed for precise estimation.

Finally we mention an interesting problem (not on SAE though) considered by Pfeffermann et al. (1998) on multi-level modeling (such as the mixed linear model (2.1)) for survey data for estimating fixed effects (β) and variance components (σ_η^2 , σ_ε^2). Here we don't have the problem of small area estimation, and the random effects η_i are defined at the PSU-level which is lower than the area level. Under a frequentist approach, they proposed a probability-weighted iterative GLS for estimating all the fixed parameters which requires knowledge of both first-stage (π_i) and second stage ($\pi_{j|i}$) selection probabilities separately, and a large number of PSUs as well as a large number of second stage units within each PSU to ensure consistency of the variance component estimates. In practice, since it is not realistic to assume large second stage sample sizes, the authors proposed scaling the weights as an option to reduce small sample bias. For a Bayesian approach as an alternative, if second order inclusion probabilities were known, it would be fairly straightforward to construct EFs for $\beta, \sigma_\eta^2, \sigma_\varepsilon^2$, and then the method of EFGL could be used to produce HB-SAE for these parameters. However, if only first order inclusion probabilities are known, as is often the case, we need to modify the EFGL method. In its present form it doesn't seem applicable, because most PSUs need to be manifested in order to have a reasonable shrinkage as mentioned earlier in Section 6. A way to modify EFGL would be to include an additional EF of the form ($\sum_{i=1}^{M} 1_{i\in s}\eta_i^2 / \pi_i - \sum_{i=1}^{M} \eta_i^2$) to account for the first stage of selection of PSU-level random effects in estimating σ_η^2, and to allow for collapsing of PSUs, if necessary, for Gaussian approximation of EFs. Note that under the usual with-replacement assumption of PSUs, design-based variances of PSU-level EFs can be estimated within each design stratum provided there are at least two PSUs per stratum.

ACKNOWLEDGMENTS

This research was supported in part by grants from U.S. Census Bureau (through NSF), from NCHS of CDC to the University of North Carolina at Chapel Hill, and from the Behavioral Sciences Branch of CDC, Department of Health and Human Services) under Contract No. 200-98-0103. The first author's research was also supported in part by a grant from Natural Sciences and Engineering Research Council of Canada held at Carleton University, Ottawa under an adjunct research professorship. Earlier versions of this paper were presented at the International Conference on Small Area Estimation and Related Topics, Potomac, MD, April 11-14, 2001, and at the FCSM conference, Arlington, VA, Nov 14-16, 2001. The authors would like to thank Jon Rao, Bob Fay, Bill Bell, Don Malec, Malay Ghosh, and Nat Schenker for their helpful and encouraging comments.

REFERENCES

Binder, D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review,* **51**, pp. 279-292.

Cassel, C.M., Särndal, C.E., and Wretman, J.H (1976), "Some Resultson Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations," *Biometrika*, **63**, pp. 615-620.

Datta, G. S. and Ghosh, M. (1991), "Bayesian Prediction in Linear Models: Applications to Small Area Estimation. *Annals of Statistics,*" **19**, pp.1746-1770.

Fay, R.E. and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein

Procedures to Census Data," *Journal of the American Statistical Association*, **74**, pp. 269-277.

Folsom, R.E., Shah, B.V. and Vaish, A. (1999), "Substance Abuse in States: A Methodological Report on Model Based Estimates from the 1994-96 NHSDAs," *Proceedings of the Survey Research Section, American Statistical Association*, pp. 371-375.

Fuller, W.A. (1975), "Regression Analysis for Sample Surveys," *Sankhya*, Ser C, **37**, pp. 117-132.

Godambe, V.P. (1960), "An Optimum Property of Regular Maximum Likelihood Estimation," *Annals of Mathematical Statistics*, **31**, pp. 1208-1212.

Godambe and Thompson, M.E, (1986), "Parameters of Super population and Survey Population, Their Relationship and Estimation," *International Statistical Review,* **54**, pp. 127-38.

Kass, R. E. and Steffey, D. (1989), "Approximate Bayesian Inferences in Conditionally Independent Hierarchical Models (Parametric Empirical Mayes Models)," *Journal of the American Statistical Association*, **84**, pp. 717-726.

Kott, P. E. (1989), "Robust Small Area Estimation Using Random Effect Modeling," *Survey Methodology*, **15**, pp. 3-12.

McCullagh, P. (1991). " Quasilikelihood and estimating functions", In *Statistical Theory and Modelling*: In honour of Sir David Cox, FRS, ed. D.V. Hinkley, N. Reid, and E.J. Snell, London: Chapman and Hall, 265-286.

Otto, M.C., and Bell, W.R. (1995), "Sampling Error Modeling of Poverty and Income Statistics for States," *Proceedings of the Government Statistics Section, American Statistical Association*, pp. 160-165.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya: The Indian Journal of Statistics*, Ser. B, **61**, 166-186.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J. (1998), "Weighting for Unequal Selection Probabilities in Multilevel Models," *Journal of the Royal Statistical Society*, B, **60**, pp. 23-40

Prasad, N.G.N. and Rao, J.N.K. (1999), “On Robust Small Area Estimates Using a Simple Random Effects Model,” *Survey Methodology*, **25**, pp. 67-72.

Singh, A.C., Stukel, D.M. and Pfeffermann, D. (1998), “Bayesian versus Frequentist Measures of Error in Small Area Estimation,” *Journal of the Royal Statistical Society*, B, **60**, pp. 377-396.

Singh, A. C. and Folsom, R. E. (2001), “Benchmarking of Small Area Estimators in a Bayesian Framework,” *International Conference on Small Area Estimation and Related Topics*, Potomac, MD April 11-14.

U.S. Census Bureau (1998), “1993 Small Area Income and Poverty Estimates (SAIPE); Small Area Estimates of School-Age Children in Poverty,” *Interim Report 2*, National Academy Press.

You, Y., and Rao, J.N.K. (2003), “Pseudo Hierarchical Bayes Small Area Estimation Combining Unit Level Models and Survey Weights,” *Journal of Statistical Planning and Inference,11,* 197-208.

Table 1: Average Posterior Mean and Standard Deviation for Model Parameters: Ignorable Sample Design

Parameter (True Value)	Average Posterior Mean				Average Posterior Standard Deviation			
	FH	EFGL	PHB	Unweighted	FH	EFGL	PHB	Unweighted
β_0 (0.5)	0.5009	0.5020	0.5020	0.5024	0.0473	0.0461	0.0482	0.0461
β_1 (1.0)	0.9946	0.9988	0.9989	0.9983	0.1650	0.0129	0.0131	0.0121
σ_η^2 (0.2)	0.1970	0.1974	0.1981	0.1981	0.0318	0.0309	0.0303	0.0303

Table 2: 95% Coverage Probability and Ratio of Prediction Interval (PI) Widths: Ignorable Sample Design

Percentiles and Means over Small Areas	Coverage Probability				Ratio of Average PI Widths		
	FH	EFGL	PHB	Unweighted	FH/EFGL	PHB/EFGL	Unweighted/EFGL
95%	0.973	0.970	0.973	0.980	1.19	1.03	1.00
75%	0.953	0.953	0.960	0.967	1.17	1.02	0.97
50%	0.940	0.940	0.953	0.953	1.16	1.01	0.91
Mean	0.942	0.941	0.950	0.950	1.16	1.01	0.89
25%	0.930	0.933	0.940	0.937	1.15	1.00	0.83
5%	0.913	0.907	0.913	0.920	1.14	1.00	0.75

Table 3: Average Posterior Mean and Standard Deviation for Model Parameters: Nonignorable Sample Design

Parameter (True Value)	Average Posterior Mean				Average Posterior Standard Deviation			
	FH	EFGL	PHB	Unweighted	FH	EFGL	PHB	Unweighted
β_0 (0.5)	0.5043	0.5029	0.5029	0.1043	0.0472	0.0450	0.0459	0.0448
β_1 (1.0)	1.0014	1.0004	1.0006	0.9999	0.1638	0.0131	0.0121	0.0103
σ_η^2 (0.2)	0.1972	0.1977	0.1909	0.1909	0.0319	0.0294	0.0290	0.0290

Table 4: 95% Coverage Probability and Ratio of Prediction Interval (PI) Widths: Nonignorable Sample Design

Percentiles and Means over Small Areas	Coverage Probability				Ratio of Average PI Widths		
	FH	EFGL	PHB	Unweighted	FH/EFGL	PHB/EFGL	Unweighted/EFGL
95%	0.973	0.970	1.000	0.007	1.91	1.54	1.35
75%	0.953	0.953	1.000	0.000	1.88	1.53	1.33
50%	0.940	0.933	0.993	0.000	1.86	1.52	1.32
Mean	0.941	0.933	0.995	0.001	1.86	1.52	1.32
25%	0.927	0.913	0.993	0.000	1.84	1.50	1.31
5%	0.910	0.897	0.987	0.000	1.82	1.49	1.30

Discussion of "Estimating Function Based Approach to Hierarchical Bayes Small Area Estimation for Survey Data"

Phillip S. Kott
National Agricultural Statistics Service

Introduction

In their intriguing paper, Singh, Folsom, and Vaish develop an Estimating-Function Hierarchical Bayesian (EFHB) methodology to replace the standard Fay-Herriot (F-H) model for small-domain estimation. I will discuss two limitations of the F-H model overcome by their EFHB methodology and two other problems that are not. This leads to the obvious question: Why combine estimating functions and hierarchical Bayesian models in the way the authors choose?

The Fay-Herriot Model

Suppose we have M small domain totals (or means) satisfying the model:

$$Y_{i+} = \mathbf{X}_{i+}\boldsymbol{\beta} + \eta_i , \qquad \eta_i \sim N(0, \sigma_\eta^2).$$

Suppose further that each of the component of the row vector $\mathbf{X}_{i+}$ is known, but each of the Y_{i+} has a randomization-based estimator:

$$y_{i+(RB)} = Y_{i+} + d_i , \qquad d_i \sim N(0, V_i) \text{ approximately.}$$

A better estimator for Y_{i+} is

$$y_{i+(\lambda)} = (1 - \lambda)y_{i+} + \lambda\, \mathbf{X}_{i+}\, \mathbf{b},$$

where $\mathbf{b}$ is an unbiased estimator for $\boldsymbol{\beta}$, $\lambda = v_i/(v_i + s_\eta^2)$ when M is large, v_i is a randomization-based estimator for V_i , and s_η^2 is an estimator for σ_η^2. A nice property of $y_{i+(\lambda)}$ is that as the sample size within domain i increases, so that V_i (and v_i) tends towards 0 under mild conditions, $y_{i+(\lambda)}$ approaches $y_{i+(RB)}$. Consequently, if $y_{i+(RB)}$ is *randomization consistent* (approaches Y_{i+} as the sample size within i grows arbitrarily large), then so is $y_{i+(\lambda)}$.

The Estimating-Function Hierarchical Bayesian Methodology

Let j be a unit within area i. The authors' EFHB technology lets us expand the F-H area-level model to the unit-level:

$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \eta_i + \varepsilon_{ij}$, where $E(\varepsilon_{ij}) = 0$.

The model holds for the population, but not necessarily for the sample. In other words, the design may be informative. In this, the authors part company from most of the small-domain literature.

If $\mathbf{X}_{i+} = \sum_{j\in U(i)} \mathbf{x}_{ij}$ is known, and $\mathbf{x}_{ij}$ is not constant within each area, then the EFHB technology produces a better estimator than $y_{i+(\lambda;F\text{-}H)}$ *under the model*. Moreover, as V_i (and v_i) approaches 0, $y_{i+(HB)}$ approaches Y_{i+}

In addition, the EFHB technology allows models of the form:

$y_{ij} = \mu(\mathbf{X}_i\boldsymbol{\beta} + \eta_i) + \varepsilon_{ij}$,

where $\mu(\,.\,)$ need not be the identity. This is particularly helpful when y_{ij} is a 0/1 variable. In that situation, $\mu(\,.\,)$ can be logistic. Unfortunately, there is a limited ability to replace $\mathbf{X}_{i+}$ with x_{ij}

The EFHB methodology uses randomization-based estimators for V_i, but such estimators are notoriously error-prone when based on small samples. Collapsing domains won't help when the estimator, v_i, is zero but V_i is positive.

Another problem with the authors' EFHB methodology is that it is not s*elf-benchmarking*. A methodology having this property produces domain estimators satisfying

$$\sum_{i=1}^{M} y_{i+} = \sum y_{i+(RB)} = y_{++(RB)} ,$$

where $y_{++(RB)}$ is model free yet has a small variance. It should be noted that the standard F-H approach is likewise *not* self-benchmarking.

Why Estimating Functions?

Arguably the first model-assisted paper (Godambe 1955) requires the estimator to be randomization unbiased. The probability-weighted ratio (and regression) estimator can have good model-based properties but has a potential randomization bias. The correct way, in my view, to deal with the randomization bias of the certain probability-weighted estimators is to change the requirement from randomization unbiasedness to randomization consistency (Isaki and Fuller 1982), which assures that the estimator in question be close to what it estimates almost surely when the sample is large enough. The wrong way is to observe that an estimator like the probability-weighted ratio can be derived from the solution to an unbiased estimating equation (e.g., Godambe 1960, Godambe &

Thompson, 1986). This "wrong way" uncovered a technique that has found many practical uses, however. It is a useful technique built on a dubious theory.

Singh, Folsom, and Vaish use estimating functions (a mild generalization of estimating equations) to generate estimators that are randomization-consistent, but not self-benchmarking. With an alternative approach, You and Rao (2003) use estimating functions to produce estimators that are both randomization-consistent and self-benchmarking. They do this by modeling the sampling variance under an ignorable model. Unfortunately, they assume a linear μ(.).

My Bottom Line

The less data we have the more we need models. Models with pre-determined functional forms have more power than semi-parametric models. Furthermore, hierarchical Bayesian models allow μ(.) to be nonlinear.

Combining estimating functions Bayesian models appear to give us the best of both worlds, the robustness of estimating functions and the power of Bayes. The former's reliance on the asymptotic normality of probability-weighted estimators, however, undercuts the advantage of a latter. We also need to ask whether sampling weights are needed because:

1. The model is correct in the population but not necessarily correct in the sample OR
2. The model may be wrong in both the sample and the population.

By positing the first, which is what the authors do, $E_{\text{model}}([y_{i+(RB)} - Y_i]^2 \mid \text{sample})$ cannot be estimated directly. Instead, one invokes the equality,

$$E_{\text{model}}\{E_{\text{rand}}([y_{i+(RB)} - Y_i]^2)\} = E_{\text{rand}}\{E_{\text{model}}([y_{i+(RB)} - Y_i]^2 \mid \text{sample})\},$$

and estimates the randomization variance for domain i, $V_i = E_{\text{rand}}([y_{i+(RB)} - Y_i]^2)$. This is often not a trivial thing to do well even with largish samples. In my view, it is much more sensible to accept the second position. Using sampling weights provides some asymptotic protection against the model being wrong in the population itself. Nevertheless, model-based parameters and predictors should be estimated as if the model were correct *and* the design noninformative. One can then estimate $E_{\text{model}}([y_{i+(RB)} - Y_i]^2 \mid \text{sample}) = E_{\text{model}}([y_{i+(RB)} - Y_i]^2)$ with relative ease, and the resulting estimator will usually have much more power than a randomization-based estimator for V_i. This is the approach effectively taken by You and Rao.

Singh, Folsom, and Vaish's EFHB approach allows them to incorporate a nonlinear μ(.) into their model. I wonder if that is enough to justify their having to rely on estimated randomization variances and put up with the inconvenience of domain estimators that are not self-benchmarking.

Session 5

Small Area and Longitudinal Estimation Using Information from Multiple Surveys

Small Area and Longitudinal Estimation Using Information from Multiple Surveys

Sharon L. Lohr
Department of Mathematics and Statistics, Arizona State University

1. Introduction

Most large sample surveys conducted by agencies such as the U.S. Bureau of the Census provide accurate statistics at the national level. Many policymakers and researchers, however, also want to obtain statistics for smaller domains such as states, counties, school districts, or demographic subgroups of a population. These domains are called small areas—so called because the sample size in the area or domain from the survey is small. The goal is to estimate θ_i, the mean value (or other characteristic) of a variable of interest y in small area i, for some or all of the small areas.

Small area estimates of income and poverty are employed in the allocation of more than eight billion dollars each year in the U.S. In that setting, no single source of information currently being collected is capable of producing reliable estimates of the number of poor people under age 18 in each county, or the number of poor children in each school district. Thus, the current practice to estimate poverty at the state level (see National Research Council, 2000, p. 49) uses auxiliary information from tax returns, food stamp programs, and the decennial census to supplement the data from the Current Population Survey (CPS). The model used is based on that in the pioneering paper by Fay and Herriot (1979). Let θ_i be the proportion of school-age children who are poor in state i. The direct estimate $\bar{y}_i$ of θ_i is calculated using data exclusively from the CPS, and $\hat{V}(\bar{y}_i)$ is an estimate of the variance of $\bar{y}_i$. A regression model for predicting θ_i using auxiliary information is

$$\theta_i = \alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ji} + v_i \tag{1}$$

where the x_{ji}'s represent covariates for state i (e.g., x_{2i} is the proportion of people receiving food stamps in state i) and v_i (assumed to follow a $N(0, \sigma_v^2)$ distribution) is the model error for state i. The regression parameters and σ_v^2 may be estimated using maximum likelihood. The predicted value from the regression equation for state i is combined with the direct estimate $\bar{y}_i$ from the CPS according to the relative amounts of information present in each estimate:

$$\hat{\theta}_i = \hat{\gamma}_i \bar{y}_i + (1 - \hat{\gamma}_i)(\hat{\alpha}_0 + \sum_{j=1}^{k} \hat{\alpha}_j x_{ji}), \tag{2}$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2 / [\hat{\sigma}_v^2 + \hat{V}(\bar{y}_i)]$. If the direct estimate is precise for a state, i.e., $\hat{V}(\bar{y}_i)$ is small, then $\hat{\gamma}_i$ is close to one and $\hat{\theta}_i$ relies mostly on the direct estimate. Conversely, if the CPS contains little information about state i's poverty rate, then $\hat{\gamma}_i$ is close to zero and $\hat{\theta}_i$ relies mostly on the predicted value from the regression. The estimator in (2) generally has smaller mean squared error (MSE) than the direct estimator $\bar{y}_i$ because it uses information available from other sources. In the extreme case where area i has no observations from the CPS and hence $\bar{y}_i$ cannot be calculated, the improvement in MSE is infinite.

Traditionally, as is done for state estimates of school-age poverty, small area estimation relies on a model relating the responses of interest in the small areas to each other and to covariates. The model allows the estimate of θ_i to "borrow strength" from other small areas through random effects terms and regression parameters. Small area estimation models have been used in many settings to obtain more accurate estimates for subpopulations without additional cost for data collection. A thorough review of research in small area estimation is given in Rao (2003).

As detailed in Rao (2003), two main types of models are used in small area estimation, distinguished by the nature of the auxiliary information. The model described above for estimating poverty rates is an example of an *area-level model*: $\bar{y}_i$, the estimate of θ_i from the survey, is related to area-level covariates. In an area-level model, the auxiliary information does not need to be known for individual persons in area i, since the covariates are summary information for the small areas. In a *unit-level model*, the response of interest for each person in area i is modeled as a function of covariates available for that person. A unit-level model might, for example, model log(income for j^{th} person in area i) using covariates of tax return and food stamp data for that person. The unit-level model thus requires that the covariate values are known (and can be linked to the income data) for the persons in the survey.

Both unit- and area-level models assume that the model covariates are measured without error. In many situations, though, auxiliary information is available that can help in the estimation, but that information is not exact. Auxiliary information may be available from another survey, or from an administrative source in which imputation has been used to fill in missing values. In both of these cases, the auxiliary information is measured with error—sampling and nonsampling error for survey data, and imputation error for incomplete administrative data. For example, the American Community Survey (ACS) will sample about 3 million households each year. For most small areas, the ACS will give relatively precise estimates of quantities it measures, and thus can be used as auxiliary information for estimating small area characteristics on many topics. The ACS still contains sampling error for many small areas, however, and that error should be included in standard errors reported for the estimates.

For another example, the U.S. National Crime Victimization Survey (NCVS) provides reliable estimates of victimization rates for the country as a whole. If separate estimates of victimization rates are desired for each state, however, some states have very small sample sizes, and standard errors using a direct estimate are unacceptably large. The same problem occurs when one desires to estimate characteristics of subgroups of the population such as victims of domestic violence—the sample sizes of domestic violence victims are not sufficiently large to give adequate precision for estimates of interest (Ybarra and Lohr, 2002). The Uniform Crime Reports (UCR), which provides statistics compiled by the FBI from law enforcement agencies, could be used as auxiliary information; Wiersema et al. (2000) found high correlations between NCVS and UCR estimates of number of victimizations using data from ten metropolitan statistical areas (MSAs). The UCR data, however, have many limitations. They only include crimes known to police; moreover, reporting is voluntary so many agencies have missing data. Even when agencies do report the data, reporting is not uniform. Maltz (1999) discussed the extent of missing data in the UCR, and described some current imputation schemes. For the UCR to be used as auxiliary information to the NCVS, imputation errors need to be incorporated into estimates of precision.

Many survey designs in the U.S. are now being integrated to allow combination of estimates. The U.S. National Health Interview Survey (NHIS) and National Health and Nutrition Examination Survey (NHANES) currently share the same primary sampling units (psu's): the psu's selected for NHIS are used as a sampling frame for NHANES. NHIS is a stratified multistage probability sample of about 100,000 persons (40,000 households) per year. The design is described in detail in Botman et al. (2000). NHANES conducts medical examinations of participants, however, and the mobile examination unit can only visit 15 psu's per year (about 5000 persons), as opposed to 358 psu's for NHIS. Because of the small sample size, NHANES data are usually accumulated over time in order to produce estimates. The small sample sizes also cause state and local estimates from NHANES to have low precision. The NHIS data provide more precise estimates of quantities measured at some localities, but the data come from an interview rather than an examination: For example, in NHANES, prevalence of diabetes may be estimated using the results of the medical exams, while in NHIS respondents are asked questions about health problems. We would expect, though, that the questionnaire results would be highly correlated with the medical examination results, and thus that the NHIS would provide high-quality auxiliary information for use with NHANES data for improved small area estimation.

The following situation is considered in this paper. Suppose there are t areas of interest (for example, $t = 50$ if states are small areas). We are interested in a characteristic θ_i of area i, for $i = 1, \ldots, t$. We have data from the primary survey for some (or all) areas, and data from an auxiliary survey for some (or all) areas. Often the characteristic of interest will be a mean or proportion. For estimating state victimization rates, θ_i might be the proportion of persons who are victims of violent crime in state i. The NCVS is considered the primary survey, and the UCR can be used to provide auxiliary information (although with error). The main questions to be considered for incorporating auxiliary information with error into small area estimates are: (1) How should the information be used in a small area model? and (2) How does the error in the auxiliary information affect the MSE of the small area estimates?

In this paper, we summarize some of our recent research on combining information from surveys to obtain more accurate estimates at the small area and national level. In Section 2, we discuss unit-level models for combining information, and in Section 3 we discuss area-level models that allow for uncertainty in the auxiliary information. Section 4 presents recent work on estimation in multiple frame surveys that can be used in small area estimation, and Section 5 discusses directions for future work.

2. Unit-level Models for Use with Multiple Surveys

Lohr and Prasad (2003) developed a framework for combining information from multiple surveys when information is available at the unit level. Let y_{ij} denote the characteristic of interest for the j^{th} unit in area i. Let $\mathbf{x}_{ij} = (x_{ij1} \cdots x_{ijk})^T$ denote a vector of other characteristics for unit j of area i. For estimating assault rates, y_{ij} might be the number of assaults that would be reported to the NCVS by person j in small area i over a specified time period, and x_{ij1} the number of assaults for that person that would be included in the UCR for the same time period. For estimating income, y_{ij} might be the log of income of household j in area i (measured in the CPS), and $\mathbf{x}_{ij}$ might be related quantities asked in the ACS. In addition, there may exist various covariates a_{ijl} that come from administrative records.

The above paragraph described applications in which y and $\mathbf{x}$ are measured from different surveys. However, the methods also apply to the "sampling on two occasions" setting. Many surveys such as the NCVS have a panel design in which the same households are sampled during several administrations of the survey. In this setting, y may be taken as the value of a characteristic on the second occasion and the auxiliary variable $\mathbf{x}$ is the same variable for the first occasion.

In area i, both $\mathbf{x}$ and y are measured on the n_i^{xy} units in $\mathcal{S}_{ixy}$; $\mathbf{x}$ (but not y) is measured on the n_i^x units in the set $\mathcal{S}_{ix}$; y (but not $\mathbf{x}$) is measured on the n_i^y units in the set $\mathcal{S}_{iy}$. If unit (ij) in the population is included in both surveys, $m = k + 1$ measurements are recorded.

We use a multivariate mixed model to describe the relationship between $\mathbf{x}$, y, and covariates. We assume that observations in different small areas are independent. To simplify expression of results, we assume that the multivariate response vector $\mathbf{u}_i$ is arranged with all observations from $\mathcal{S}_{ixy}$ first, followed by those from $\mathcal{S}_{ix}$ and $\mathcal{S}_{iy}$, so

$$\mathbf{u}_i^T = [\mathbf{x}_{i1}^T, y_{i1}, \ldots, \mathbf{x}_{i,n_i^{xy}}^T, y_{i,n_i^{xy}}, \mathbf{x}_{i,n_i^{xy}+1}^T, \ldots, \mathbf{x}_{i,n_i^{xy}+n_i^x}^T, y_{i,n_i^{xy}+n_i^x+1}, \ldots, y_{i,n_i^{xy}+n_i^x+n_i^y}].$$

Let

$$\mathbf{u}_i = \mathbf{A}_i\boldsymbol{\mu} + \mathbf{Z}_i\mathbf{v}_i + \mathbf{e}_i \tag{3}$$

where $\boldsymbol{\mu}$ is a vector of fixed effects parameters, $\mathbf{A}_i$ and $\mathbf{Z}_i$ are known matrices, and $\mathbf{v}_i$ and $\mathbf{e}_i$ are independent random vectors with mean $\mathbf{0}$. $\text{Cov}\,(\mathbf{v}_i) = \boldsymbol{\Sigma}_v$ and

$$\text{Cov}\,(\mathbf{e}_i) = \mathbf{R}_i = [\mathbf{I}_{n_i^{xy}} \bigotimes \boldsymbol{\Sigma}_e] \bigoplus [\mathbf{I}_{n_i^x} \bigotimes \boldsymbol{\Sigma}_{exx}] \bigoplus [\mathbf{I}_{n_i^y} \bigotimes \boldsymbol{\Sigma}_{eyy}],$$

where the matrices $\boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_e$ are partitioned as

$$\boldsymbol{\Sigma}_v = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{vxx} & \boldsymbol{\Sigma}_{vxy} \\ \boldsymbol{\Sigma}_{vxy}^T & \Sigma_{vyy} \end{array} \right], \quad \boldsymbol{\Sigma}_e = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{exx} & \boldsymbol{\Sigma}_{exy} \\ \boldsymbol{\Sigma}_{exy}^T & \boldsymbol{\Sigma}_{eyy} \end{array} \right]$$

and where $\bigoplus$ represents direct sum and $\bigotimes$ represents Kronecker product. Thus

$$\mathbf{V}_i = \text{Cov}\,(\mathbf{u}_i) = \mathbf{R}_i + \mathbf{Z}_i\boldsymbol{\Sigma}_v\mathbf{Z}_i^T \tag{4}$$

with

$$\mathbf{Z}_i = \left[\begin{array}{lcl} \mathbf{1}_{n_i^{xy}} & \otimes & \mathbf{I}_m \\ \mathbf{1}_{n_i^x} & \otimes & (\mathbf{I}_k \; \mathbf{0}_k) \\ \mathbf{1}_{n_i^y} & \otimes & (\mathbf{0}_k^T \; 1) \end{array}\right]$$

where $\mathbf{1}_j$ is a j-vector of ones.

For simplicity of presentation, we take $\boldsymbol{\mu}$ to be the m-vector of fixed effects means, partitioned as $\boldsymbol{\mu}^T = [\boldsymbol{\mu}_x^T \; \mu_y]$. However, all results are easily extended to the case where $\boldsymbol{\mu}$ is a general vector of parameters, and $\mathbf{A}_i$ is a matrix of fixed effects covariates. In this way information from a census or from administrative records may be incorporated into the small area estimates through regression.

Under this setup, Lohr and Prasad (2003) showed that if $\boldsymbol{\mu}$ and the covariance component matrices are known, then the best linear unbiased predictor (BLUP) for $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{ix}^T, \theta_i)^T$ is

$$\tilde{\boldsymbol{\mu}}_i = \boldsymbol{\mu} + n_i^{xy}\mathbf{M}_i\boldsymbol{\Sigma}_e^{-1}(\bar{\mathbf{u}}_{ixy} - \boldsymbol{\mu}) + \mathbf{M}_i\mathbf{n}_i^*(\boldsymbol{\Sigma}_e^*)^{-1}(\bar{\mathbf{u}}_i^* - \boldsymbol{\mu}). \tag{5}$$

Here, $\bar{\mathbf{u}}_{ixy}$ is the average of the n_i^{xy} vectors $(\mathbf{x}_{ij}^T, y_{ij})^T$ for $j \in \mathcal{S}_{ixy}$; $\bar{\mathbf{u}}_i^* = (\bar{\mathbf{x}}_{ix}^T, \bar{y}_{iy})^T$ contains the averages of the $\mathbf{x}_{ij}$'s for $j \in \mathcal{S}_{ix}$ and of the y_{ij}'s for $j \in \mathcal{S}_{iy}$;

$$\boldsymbol{\Sigma}_e^* = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{exx} & 0 \\ 0 & \boldsymbol{\Sigma}_{eyy} \end{array}\right], \tag{6}$$

$$\mathbf{n}_i^* = \left[\begin{array}{cc} n_i^x\mathbf{I} & 0 \\ 0 & n_i^y \end{array}\right], \tag{7}$$

and

$$\mathbf{M}_i = (\boldsymbol{\Sigma}_v{}^{-1} + n_i^{xy}\boldsymbol{\Sigma}_e^{-1} + \mathbf{n}_i^*(\boldsymbol{\Sigma}_e^*)^{-1})^{-1}. \tag{8}$$

This estimator reduces to the multivariate estimator in Datta et al. (1999) if $n_i^x = n_i^y = 0$.

The BLUP $\tilde{\theta}_i$ for θ_i is the m^{th} component of $\tilde{\boldsymbol{\mu}}_i$, and $\text{MSE}(\tilde{\theta}_i) = \mathbf{M}_{iyy}$, the (m, m) entry of $\mathbf{M}_i$. As a special case, the BLUP of θ_i when $n_i^{xy} = n_i^y = 0$ is $\tilde{\theta}_i = \mu_y + \boldsymbol{\Sigma}_{vxy}^T\boldsymbol{\Sigma}_{vxx}^{-1}(\tilde{\boldsymbol{\mu}}_{ix} - \boldsymbol{\mu}_x)$: the estimator then borrows strength by using the between-area covariance of $\mathbf{x}$ and y.

If the quantities from the two surveys are correlated, $\tilde{\theta}_i$ is more efficient than the corresponding estimator that does not use the auxiliary survey data. Lohr and Prasad (2003) derived the gain in efficiency, and showed that $\tilde{\theta}_i$ has smaller MSE than the estimator from the univariate unit-level model of Battese et al. (1988) if $n_i^x n_i^{xy}\boldsymbol{\Sigma}_{exy} \neq 0$ or $n_i^x\boldsymbol{\Sigma}_{vxy} \neq 0$.

2.1. Estimation of Unknown Quantities

The estimator in (5) was calculated assuming that the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_v$, and $\boldsymbol{\Sigma}_e$ are known. In practice, these must be estimated from the data.

Using the generalized least squares estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$, and using consistent estimators of the covariance components, the multivariate estimator becomes

$$\hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{\mu}} + n_i^{xy}\hat{\mathbf{M}}_i\hat{\boldsymbol{\Sigma}}_e^{-1}(\bar{\mathbf{u}}_{ixy} - \hat{\boldsymbol{\mu}}) + \hat{\mathbf{M}}_i\mathbf{n}_i^*(\hat{\boldsymbol{\Sigma}}_e^*)^{-1}(\bar{\mathbf{u}}_i^* - \hat{\boldsymbol{\mu}}). \tag{9}$$

Lohr and Prasad (2003) derived the second order asymptotic properties of this estimator. As with the BLUP, $\hat{\theta}_i$ is the m^{th} component of $\hat{\boldsymbol{\mu}}_i$.

The method has been implemented in R and S-Plus, with restricted maximum likelihood used to estimate the covariance components. A simulation study demonstrated that $\hat{\theta}_i$ was much more efficient than an estimator that did not use the information from the auxiliary survey, particularly when $\boldsymbol{\Sigma}_{vxy}$ was large relative to $\boldsymbol{\Sigma}_{vxx}$ and Σ_{vyy}. Even with relatively modest sample sizes in the auxiliary survey (say, $n_i^x = 5$), when the survey quantities were highly correlated the MSE of $\hat{\theta}_i$ was about 1/5 of the MSE of the univariate unit-level estimator that did not use the $\mathbf{x}$ information.

When using the multivariate estimator with separate surveys, in most cases it will not be necessary to match sample observations between the two surveys. Even when the survey designs share the same primary sampling units, it is unlikely that the same persons are included in the surveys. Thus, it is overwhelmingly probable that in most small areas, $n_i^{xy} = 0$. Consequently, the estimator in (5) will involve $\boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_e^*$ but not $\boldsymbol{\Sigma}_{exy}$. The vector $\boldsymbol{\Sigma}_{exy}$ is the only quantity, however, whose estimation requires that units in the two surveys be matched. The matrix $\boldsymbol{\Sigma}_e^*$ can be estimated from the two separate surveys, and $\boldsymbol{\Sigma}_v$ can be estimated provided that the number of small areas that contain observations from both surveys is sufficiently large.

2.2. Robust Estimation of Covariance Components

The unit-level multivariate approach depends on a model, and the estimates are therefore sensitive to departures from that model. The estimates of the fixed effects and of the covariance components can perform badly in the presence of aberrant observations. In particular, the restricted maximum likelihood estimates of the covariance components that were used in (9) are affected by outliers. Outliers will not be too great of a problem for estimating $\boldsymbol{\Sigma}_e$ because in most situations there will be sufficient degrees of freedom at the within-area level to mitigate the effect of a few moderate outliers. There are fewer degrees of freedom for estimating $\boldsymbol{\Sigma}_v$, however, so if the estimated mean of a small area is aberrant, this outlying area may greatly affect the REML estimate of $\boldsymbol{\Sigma}_v$.

Dueck and Lohr (2003) developed a method for robust estimation of multivariate covariance components. They used multivariate M-estimation of random effects to reduce the influence of outliers—at both the within-area and between-area levels—on the estimated covariance components. Preliminary research indicates that use of this method, together with robust estimation of the fixed effects, improves the accuracy of small area estimates when some data may be contaminated.

3. Area-level Models for Multiple Surveys

The models in Section 2 result in improved efficiency when unit-level auxiliary information exists and observations can be matched across surveys. Matching is easy when sampling on two occasions, where y is the response of interest measured at time 2 and the auxiliary information is the same response measured at time 1. In other settings, only areas may need to be matched, since different units will be used in the two surveys. For some applications, however, matching units may be infeasible: records from the NCVS cannot in general be matched with the same persons' records from the UCR. In addition, there may be concerns that using unit-level data across surveys or other data sources may compromise confidentiality of the data (see Lohr, 2003). For some surveys, respondents may not have given permission to have their data combined with individual-level information from other sources. In such cases, area-level models are preferred.

In this section, we examine area-level models for use with two surveys. To simplify presentation, we consider the case where θ_i is a population mean, although extensions to other parameters are readily made. Let $\bar{y}_i$ be an unbiased estimator of θ_i from the primary survey, with sampling variance $V(\bar{y}_i) = \psi_i$. Administrative data for area i, $\mathbf{A}_i$, is assumed to be measured without error. We consider the k-vector $\mathbf{X}_i$ to be population characteristics for area i which in some areas can be estimated by a vector $\mathbf{x}_i$ from the auxiliary data source. Often, $\mathbf{X}_i$ will be a vector of population means for area i. We assume here that when $\mathbf{x}_i$ is measured, $E(\mathbf{x}_i) = \mathbf{X}_i$ and $V(\mathbf{x}_i) = \mathbf{\Sigma}_i$.

3.1. What if Error in Auxiliary Information is Ignored?

The Fay-Herriot (1979) model leads to the BLUP of θ_i. If $\bar{y}_i$ and θ_i are assumed to be normally distributed, the Fay-Herriot estimator can be motivated in an empirical Bayesian framework (see Rao, 2003, chapter 9). It is assumed that $\bar{y}_i \mid \theta_i, \psi_i \sim N(\theta_i, \psi_i)$; a regression model for the population quantity is given by

$$\theta_i|\mathbf{A}_i, \mathbf{X}_i, \sigma_v^2, \boldsymbol{\alpha}, \boldsymbol{\beta} \sim N(\mathbf{A}_i^T\boldsymbol{\alpha} + \mathbf{X}_i^T\boldsymbol{\beta}, \sigma_v^2). \tag{10}$$

If the quantities $(\bar{y}_i, \theta_i)$ are independent for $i = 1, \ldots, t$, then the posterior distribution of θ_i is

$$\theta_i|\bar{y}_i, \mathbf{A}_i, \mathbf{X}_i, \sigma_v^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \psi_i \sim N[\gamma_i^*\bar{y}_i + (1-\gamma_i^*)(\mathbf{A}_i^T\boldsymbol{\alpha} + \mathbf{X}_i^T\boldsymbol{\beta}), \psi_i\gamma_i^*] \tag{11}$$

where $\gamma_i^* = \sigma_v^2/(\sigma_v^2 + \psi_i)$. The mean of the posterior distribution of θ_i is

$$\tilde{\theta}_{iEB} = \gamma_i^*\bar{y}_i + (1-\gamma_i^*)(\mathbf{A}_i^T\boldsymbol{\alpha} + \mathbf{X}_i^T\boldsymbol{\beta}). \tag{12}$$

Now let us examine what happens if an estimator $\hat{\mathbf{X}}_i$ with $\text{MSE}(\hat{\mathbf{X}}_i) = \mathbf{C}_i$ is substituted for the population quantity $\mathbf{X}_i$ in (12); either $\mathbf{x}_i$ or another estimator may be used for $\hat{\mathbf{X}}_i$. Let

$$\tilde{\theta}_i^* = \gamma_i^*\bar{y}_i + (1-\gamma_i^*)(\mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta}). \tag{13}$$

Then $\text{MSE}(\tilde{\theta}_i^*) = \psi_i\gamma_i^* + (1-\gamma_i^*)^2\boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta}$, and the posterior variance in (11) underestimates the true variance. If the matrix $\mathbf{C}_i$ is large, the mean squared error of the $\tilde{\theta}_i^*$ can be larger than ψ_i, so that the supposedly improved small area estimator can perform worse than the direct estimator that uses no auxiliary information. In addition, if the error in estimating $\mathbf{X}_i$ is ignored and $\psi_i\gamma_i^*$ is naively reported to be the MSE, the estimator will be thought to be more precise than it really is.

We can correct the MSE by incorporating the error in estimating $\mathbf{X}_i$ into the model in (10). If $\mathbf{x}_i|(\mathbf{X}_i, \mathbf{\Sigma}_i) \sim N(\mathbf{X}_i, \mathbf{\Sigma}_i)$, $\mathbf{X}_i|(\boldsymbol{\mu}_x, \mathbf{\Sigma}) \sim N(\boldsymbol{\mu}_x, \mathbf{\Sigma})$, and the quantities $(\mathbf{x}_i, \mathbf{X}_i)$ are independent across areas, then the posterior distribution of θ_i has mean

$$\gamma_i^* y_i + (1 - \gamma_i^*)(\mathbf{A}_i^T \boldsymbol{\alpha} + \mathbf{c}_i^T \boldsymbol{\beta})$$

and variance

$$\psi_i \gamma_i^* + (1 - \gamma_i^*)^2 \boldsymbol{\beta}^T \mathbf{D}_i \boldsymbol{\beta},$$

where

$$\mathbf{c}_i = \mathbf{\Sigma}(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\mathbf{x}_i + \mathbf{\Sigma}_i(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\boldsymbol{\mu}_x$$

and

$$\mathbf{D}_i = \mathbf{\Sigma}(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\mathbf{\Sigma}_i = \mathbf{\Sigma}_i(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\mathbf{\Sigma}.$$

With the additional assumptions on the distribution of the auxiliary survey data, the posterior variance is correct for the MSE. The relative weight γ_i^*, however, still does not account for the error in estimating $\mathbf{X}_i$; it is possible for the posterior variance to be larger than ψ_i so that incorporating the auxiliary $\mathbf{x}$ information may result in a decrease in precision. The methods in the following sections use the uncertainty about $\mathbf{x}_i$ when determining the relative weightings of the direct and indirect estimators.

3.2. Multivariate Fay-Herriot Model

Fay (1987) and Datta et al. (1991) developed a Fay-Herriot-type model for a multivariate response, and showed that it often results in more efficient estimators for a small area quantity of interest than the univariate Fay-Herriot model. Datta et al. (1991) were interested in estimating the median income of four-person households in state i. The direct estimate was from the CPS. The auxiliary information, $x_i = (3/4)$ (median income of five-person households) + (1/4) (median income of three-person households) also came from the CPS. The multivariate model they used reduced the MSE of the estimator of θ_i through correlations with the other variables. Lohr and Ybarra (2003) extended this model to allow for missing observations, and to allow the observations to come from different sources. The following summarizes the results for the notationally simpler case when $\mathbf{x}_i$ and $\bar{y}_i$ are independent.

Let $\mathbf{U}_i = [\mathbf{X}_i^T,\ \theta_i]^T$ represent the population values for each of the i areas, $i = 1, \ldots, t$. Define $\mathbf{T}_i$ to be the matrix whose j^{th} row is $[\mathbf{0}^T, \cdots, \mathbf{0}^T, \mathbf{A}_i^T, \mathbf{0}^T, \cdots, \mathbf{0}^T]$ where the $\mathbf{A}_i^T$ occurs as the j^{th} column. Consider the model

$$\mathbf{U}_i = \mathbf{T}_i \boldsymbol{\alpha} + \mathbf{v}_i \tag{14}$$

where $\mathbf{v}_i \sim N(\mathbf{0}, \mathbf{\Sigma}_b)$ and $\boldsymbol{\alpha}$ is a vector of regression coefficients. As in the unit-level model, the covariance matrix $\mathbf{\Sigma}_b$ is partitioned as

$$\mathbf{\Sigma}_b = \begin{bmatrix} \mathbf{\Sigma}_{bxx} & \mathbf{\Sigma}_{bxy} \\ \mathbf{\Sigma}_{bxy}^T & \mathbf{\Sigma}_{byy} \end{bmatrix}.$$

Define the vector $\mathbf{u}_i$ and the matrices $\mathbf{Z}_i$ and $\mathbf{\Psi}_i$ for three cases:

1. If $\mathbf{x}$ and y are both observed for area i then $\mathbf{u}_i = [\mathbf{x}_i^T, \bar{y}_i]^T$, $\mathbf{Z}_i = \mathbf{I}_{k+1}$, and
$$\boldsymbol{\Psi}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{0} \\ \mathbf{0}^T & \psi_i \end{bmatrix}.$$
2. If $\mathbf{x}$ is observed in area i but y is not observed then $\mathbf{u}_i = \mathbf{x}_i$, $\mathbf{Z}_i^T = [\mathbf{I}_k, \mathbf{0}_k]$ and $\boldsymbol{\Psi}_i = \boldsymbol{\Sigma}_i$
3. If y is observed in area i but $\mathbf{x}$ is not observed then $\mathbf{u}_i = \bar{y}_i$, $\mathbf{Z}_i^T = [\mathbf{0}_k^T, 1]$, and $\boldsymbol{\Psi}_i = \psi_i$.

Then the observations $\mathbf{u}_i$ follow the model
$$\mathbf{u}_i = \mathbf{Z}_i^T \mathbf{T}_i \boldsymbol{\alpha} + \mathbf{Z}_i^T \mathbf{v}_i + \mathbf{e}_i, \tag{15}$$
where $\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Psi}_i)$. The covariance matrix of $\mathbf{u}_i$ is
$$\mathbf{V}_i = \mathbf{V}_i(\mathbf{u}_i) = \mathbf{Z}_i^T \boldsymbol{\Sigma}_b \mathbf{Z}_i + \boldsymbol{\Psi}_i.$$
The $\mathbf{u}_i$'s are assumed to be independent. This model then fits into the block diagonal covariance structure model described in Section 6.3 of Rao (2003). Define
$$\tilde{\boldsymbol{\alpha}} = \left(\sum_i \mathbf{T}_i^T \mathbf{Z}_i \mathbf{V}_i^{-1} \mathbf{Z}_i^T \mathbf{T}_i \right)^{-1} \left(\sum_i \mathbf{T}_i^{-1} \mathbf{Z}_i \mathbf{V}_i^{-1} \mathbf{u}_i \right),$$
$$\mathbf{K}_i = (\boldsymbol{\Sigma}_{bxx} + \boldsymbol{\Sigma}_i)^{-1},$$
and
$$\kappa_i = \frac{\Sigma_{byy} - \boldsymbol{\Sigma}_{bxy}^T \mathbf{K}_i \boldsymbol{\Sigma}_{bxy}}{\Sigma_{byy} - \boldsymbol{\Sigma}_{bxy}^T \mathbf{K}_i \boldsymbol{\Sigma}_{bxy} + \psi_i},$$
The BLUP for $(\mathbf{X}_i^T, \theta_i)$ is then
$$\tilde{\theta}_{iMFH} = \kappa_i \bar{y}_i + (1 - \kappa_i) \left[[\mathbf{0}^T, 1] \mathbf{T}_i \tilde{\boldsymbol{\alpha}} + \boldsymbol{\Sigma}_{bxy}^T \mathbf{K}_i \left(\mathbf{x}_i - [\mathbf{I}, \mathbf{0}] \mathbf{T}_i \tilde{\boldsymbol{\alpha}} \right) \right] \tag{16}$$
if both $\mathbf{x}_i$ and $\bar{y}_i$ are observed in area i;
$$\tilde{\theta}_{iMFH} = [\mathbf{0}^T, 1] \mathbf{T}_i \tilde{\boldsymbol{\alpha}} + \boldsymbol{\Sigma}_{bxy}^T \mathbf{K}_i (\mathbf{x}_i - [\mathbf{I}, \mathbf{0}] \mathbf{T}_i \tilde{\boldsymbol{\alpha}}) \tag{17}$$
if $\mathbf{x}_i$ is observed in area i but $\bar{y}_i$ is not;
$$\tilde{\theta}_{iMFH} = \kappa_i \bar{y}_i + (1 - \kappa_i) [\mathbf{0}^T, 1] \mathbf{T}_i \tilde{\boldsymbol{\alpha}} \tag{18}$$
if $\bar{y}_i$ is observed but $\mathbf{x}_i$ is not.

The weighting κ_i in the small area estimator in (16) to (18) depends on the variability of $\mathbf{x}_i$ as well as on the sampling variability of $\bar{y}_i$: κ_i is smaller, and the small area estimator depends more heavily on the direct estimator, if the variability of $\mathbf{x}_i$ is larger. If $\mathbf{X}_i$ is measured exactly (i.e., all entries of $\boldsymbol{\Sigma}_i$ are 0), then $\tilde{\theta}_{iMFH}$, using assumptions of normality, coincides with the univariate Fay-Herriot estimator that incorporates the $\mathbf{X}_i$'s as covariates.

The MSE of the estimator in (16) to (18) can be obtained using standard methods and is given in Lohr and Ybarra (2003). As occurred with the unit-level model, use of the multivariate Fay-Herriot model results in improved efficiency.

In practice, $\boldsymbol{\Sigma}_b$ as well as $\boldsymbol{\alpha}$ must be estimated from the data. Method of moments, maximum likelihood, or restricted maximum likelihood may be used. See Datta et al. (2001) for a comparison of the estimators of $\boldsymbol{\Sigma}_b$ in the univariate case.

3.3. Measurement Error Model

As shown in Section 3.1, ignoring the error in $\mathbf{x}_i$ gives a biased mean squared error and a non-optimal weighting of the direct and indirect estimators. The motivation for using a measurement error model comes from the observation that omitted or inaccurate covariates can cause bias. Suppose that the model in (10) holds, but it is fitted omitting the term $\mathbf{X}_i^T\boldsymbol{\beta}$. Then estimates of the regression parameters $\boldsymbol{\alpha}$ and the predicted values may be biased. This bias leads to an increase in the MSE of the predicted values. If $\mathbf{x}_i$ or another estimator $\hat{\mathbf{X}}_i$ is included in the covariates, however, the error in measuring $\mathbf{X}_i$ must be accounted for in the estimation and mean squared error. Fuller (1987, 1990), Carroll et al. (1995) and Cheng and Van Ness (1999) discussed measurement error models for estimation of regression parameters and for prediction.

As before, let $\hat{\mathbf{X}}_i$ be an estimator of the population quantity $\mathbf{X}_i$ with MSE$\,(\hat{\mathbf{X}}_i) = \mathbf{C}_i$. We assume that such an estimator exists for every area: If $\mathbf{x}$ is not measured in area i, then an empirical Bayes estimator or imputed value may be used for $\hat{\mathbf{X}}_i$. Consider the model

$$\bar{y}_i = \mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta} + r_i(\hat{\mathbf{X}}_i, \mathbf{X}_i) + e_i \tag{19}$$

where

$$r_i(\hat{\mathbf{X}}_i, \mathbf{X}_i) = v_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)^T\boldsymbol{\beta}.$$

Here, $v_i \sim N(0, \sigma_v^2)$ represents the model error and $e_i \sim N(0, \psi_i)$ represents the design-based survey error for $\bar{y}_i$. We assume that v_i is independent of both $\hat{\mathbf{X}}_i$ and $\bar{y}_i$. For simplicity, we also assume here that all $\hat{\mathbf{X}}_i$'s and $\bar{y}_i$'s are independent; Ybarra (2003) develops theory for the more general case. Consequently, MSE$(r_i) = \sigma_v^2 + \boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta}$. Now let

$$\tilde{\theta}_{iME} = \gamma_i\bar{y}_i + (1-\gamma_i)(\mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta}), \tag{20}$$

where

$$\gamma_i = \frac{\sigma_v^2 + \boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta}}{\sigma_v^2 + \boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta} + \psi_i}. \tag{21}$$

Then $\tilde{\theta}_{iME}$ has minimum mean squared error among all linear combinations of $\bar{y}_i$ and $\mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta}$ of the form $a_i\bar{y}_i + (1-a_i)(\mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta})$ where $0 \le a_i \le 1$. The estimator in (21) may also be derived as the "best" estimator in the Rao-Blackwell sense if normality is assumed.

The relative weights γ_i depend on the error in estimating $\mathbf{X}_i$: γ_i is smaller when $\hat{\mathbf{X}}_i$ is measured without error. If $\hat{\mathbf{X}}_i$ is measured imprecisely, then γ_i is larger and the estimator depends more heavily on the direct estimator $\bar{y}_i$. If $\bar{y}_i$ is measured in area i then MSE$(\tilde{\theta}_{iME}) = \psi_i\gamma_i$, which is at most as large as the variance ψ_i of the direct estimator, $\bar{y}_i$. If $\bar{y}_i$ is not measured in area i then MSE$\,(\tilde{\theta}_{iME}) = \sigma_v^2 + \boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta}$.

Note that MSE$\,(\tilde{\theta}_{iME}) \le$ MSE$\,(\tilde{\theta}_i^*)$ where $\tilde{\theta}_i^*$ is the substitution estimator from (13): the two MSE's are equal if $\mathbf{C}_i = 0$. If the empirical Bayes estimator is used for $\hat{\mathbf{X}}_i$, then it can be shown that the estimator in (20) is equivalent to the multivariate Fay-Herriot estimator.

In practice, the quantities σ_v^2, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown and must be estimated from the data. Lindley (1947, p. 243) suggested using weighted least squares to estimate the regression parameters. For our model, the MSE of the errors $(r_i + e_i)$ is $\psi_i + \sigma_v^2 + \boldsymbol{\beta}^T \mathbf{C}_i \boldsymbol{\beta}$. Thus, one can solve for the unknown parameters by minimizing

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{m} \frac{(\bar{y}_i - \mathbf{A}_i^T \boldsymbol{\alpha} - \hat{\mathbf{X}}_i^T \boldsymbol{\beta})^2}{\psi_i + \sigma_v^2 + \boldsymbol{\beta}^T \mathbf{C}_i \boldsymbol{\beta}}$$

where the sum is over areas i where $\bar{y}_i$ is measured. Gleser (1981) gave large sample properties of the resulting estimates of the regression parameters. If σ_v^2 is unknown, we can use modified least squares to estimate the parameters (Cheng and Van Ness, 1999, pp. 85 and 146). In this case an unbiased estimator of σ_v^2 is

$$Q_1(\boldsymbol{\alpha}, \boldsymbol{\beta}, \psi_1, \ldots, \psi_m) = m^{-1} \sum_{i=1}^{m} [(\bar{y}_i - \mathbf{A}_i^T \boldsymbol{\alpha} - \hat{\mathbf{X}}_i^T \boldsymbol{\beta})^2 - \psi_i - \boldsymbol{\beta}^T \mathbf{C}_i \boldsymbol{\beta}] \tag{22}$$

Minimizing Q_2 with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ gives estimates of the regression parameters. Note, though, that terms in (22) may be negative and it is possible that minimization will occur on the boundaries of the parameter space. Ybarra (2003) modified the estimators so that the expected values of the regression parameters are finite and derived properties of the models using these estimators. She also explored effects of estimating the variances from the data.

Although in some situations the measurement error model and multivariate Fay-Herriot method give similar results, we prefer the measurement error model for many practical situations. It is more flexible for choice of estimator of $\mathbf{X}_i$. In addition, robust methods may be used for estimating the regression parameters and variance terms, so that the measurement error model is adaptable for situations in which some of the $\mathbf{x}_i$'s are outliers due to variable quality of the data sources.

3.4. Applications

The measurement error model has an advantage over the multivariate Fay-Herriot approach in that means and variances of the auxiliary information can be estimated separately from the quantities from the primary survey. Missing values may be imputed, and imputation variance used for the MSE of $\hat{\mathbf{X}}_i$. This approach would work better than the multivariate Fay-Herriot approach for estimating victimization rates at the state level, using the Uniform Crime Reports (UCR) data as auxiliary information.

The UCR data sets give crimes reported each month by each of the approximately 19,000 law enforcement agencies in the United States. In a typical year, however, approximately 1/3 of the total possible month/agency cells are missing. If complete records only are used as auxiliary information in a Fay-Herriot-type model, the resulting small area estimates may be biased and will have standard errors that are too small because they do not account for the uncertainty in the auxiliary information. The multivariate Fay-Herriot approach can reduce some of this bias by incorporating administrative covariates to improve prediction of the UCR (essentially, including the imputation in the model). But the imputation will be done at the state level for annual data; this will not be as good as an imputation done separately using partial agency information and longitudinal trends with the monthly data.

Schalk (2003) studied imputation methods for the western region of the Uniform Crime Reports data. She evaluated the currently used hot deck method, nearest neighbor, and several regression models for imputing missing cells and found that the hot deck method is the least accurate. All of the models studied can give standard errors for the statewide quantities by using bootstrap or multiple imputation. Thus, by doing the imputation separately, the auxiliary information is more accurate and is accompanied by an estimate of precision $\mathbf{C}_i$ that can be used with the measurement error model for estimating victimization rates with NCVS data. With the imputed values from the UCR, we are now in a position to apply the measurement error models in Section 3.3 to obtain more accurate estimates of local victimization rates.

We are also currently using the models discussed in this section to obtain small area estimates of the prevalence of diabetes for 50 demographic subgroups based on race/ethnicity, gender, and age. In NHANES, diabetes prevalence is estimated using medical exams of plasma glucose levels, while in NHIS diabetes-related problems are assessed using the results of questionnaires. Correlation between the items in the two surveys is about 0.4; using the NHIS data as auxiliary information reduces the MSE for diabetes prevalence in small demographic groups (with NHANES sample sizes between 5 and 7) by 40-80%.

4. Multiple Frame Surveys for Small Area Estimation

Up to this point, we have discussed using a second survey to provide auxiliary information for estimating a quantity of interest measured in the primary survey. The models given in Sections 2 and 3 use all available information for predicting θ_i; if area i has no observations from either the primary or secondary survey, then $\hat{\theta}_i$ relies on the predicted value from the regression using the administrative data. This may be the best that can be done with the available information, but sometimes a different design can give more precision for the direct estimators and for the estimated regression parameters.

One such design that can be used is a multiple frame survey. In a multiple frame survey, probability samples are drawn independently from Q frames $A_1, \ldots, A_Q$. The union of the Q frames is assumed to cover the finite population of interest, $\mathcal{U}$. The frames may overlap, resulting in a possible $2^Q - 1$ nonoverlapping domains.

Rao (2003, chapter 2) discussed the use of multiple frame designs for improving small area estimation. The primary purpose of many surveys is estimation of quantities such as unemployment or criminal victimization at the national level; the designs for the surveys thus are directed toward the national estimates, even though some surveys contain design features useful for small area estimation. These surveys, though, can be supplemented with additional samples from small areas of interest, so that the original survey and additional samples can be considered as a multiple frame survey. Madans et al. (2001) discussed using multiple frame surveys for supplementing information from NHIS; additional surveys may be taken from different states and combined with NHIS data for improved estimation at the state level. In this situation the same questions may be used in NHIS and the supplementary surveys.

Various estimators that have been proposed for combining information from the separate samples were reviewed in Lohr and Rao (2000). These estimators modify the weights associated with sampled units from each frame, so that the overall population total is estimated by a weighted sum of the observations from all of the samples using the modified weights. Many of these methods, however, were developed for estimating one population total or mean at a time, and use a different set of modified weights for each characteristic of interest. Such an approach will give nearly optimal results for individual responses, but will not work well for estimating small area totals or means directly from the surveys: If different weights are used for estimating the population total in different small areas, the sum of estimated small area population totals will not equal the estimated total for larger areas. It is thus desirable to have methods for obtaining direct small area estimates from multiple frame surveys that use the same set of weights for all variables. Skinner and Rao (1996) developed a pseudo-maximum likelihood method that uses the same weights for all variables for the two-frame situation.

Lohr and Rao (2002) developed estimation methods for multiple frame surveys with more than two frames that use the same weights for each variable being estimated, and thus can be applied when supplemental surveys are taken in several small areas. These methods easily apply to the small area setting by letting the variable of interest be the value θ_i for the i^{th} small area. The improved direct estimators of the θ_i's may then be used with an area-level model to achieve greater efficiency.

5. Discussion and Future Work

In this paper, we have summarized recent research we have done on combining information from different sources for small area estimation. In many situations, much greater efficiency can be achieved by using auxiliary information from another survey. We believe that these methods have the potential to increase the accuracy of small area estimates with no or minimal increase in the cost of data collection, as they are all based on more efficient use of existing data.

The American Community Survey is intended, through its large sample size, to provide improved direct small area estimates for income and poverty. Those characteristics and other quantities measured in the ACS can also provide valuable and timely auxiliary data for small area estimation of quantities measured in other surveys. The methods summarized in this paper can be used to take advantage of this new, detailed data source for small area estimation of many different characteristics of interest.

Since the ACS uses rolling samples, longitudinal methods will also be helpful when using the ACS as auxiliary information. We are currently working on incorporating time series models into the estimation, and on obtaining longitudinal estimates from multiple frame surveys. A related problem is using spatial models to better include geographic information.

Another important problem under study is robustness to the model and to methods for estimating model quantities. One challenge of using UCR data as auxiliary information for the NCVS, in addition to the missing values, is that some agencies provide inaccurate estimation. These inaccuracies could then bias the results. Using robust methods is expected to reduce the effects of possible UCR outliers and result in more accurate small area estimates.

Acknowledgements

This research was supported by the National Science Foundation through Grant SES 0105852. We are grateful for this support, and for helpful comments from J.N.K. Rao, John Eltinge, and the late Charles Alexander.

References

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). "An Error-components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28–36.

Botman, S.L., Moore, T.F., Moriarity, C.L. and Parsons, V.L. (2000). "Design and Estimation for the National Health Interview Survey, 1995-2004," National Center for Health Statistics. *Vital Health Statistics* 2(130).

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models,* London: Chapman & Hall.

Cheng, C.-L. and Van Ness, J.W. (1999). *Statistical Regression with Measurement Error.* London: Arnold.

Datta, G. A., B. Day and I. Basawa (1999). "Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation," *Journal of Statistical Planning and Inference*, 75, 269–279.

Datta, G. S., R. E. Fay and M. Ghosh (1991), "Hierarchical and Empirical Multivariate Analysis in Small Area Estimation," *Proceedings of the Bureau of the Census Annual Research Conference*, Washington: Bureau of the Census, 63-79.

Datta, G. S., J. N. K. Rao and Smith, D. D. (2001), "On Measures of Uncertainty of Small Area Estimators in the Fay-Herriot Model," Technical Report, University of Georgia, Athens, Georgia.

Dueck, A. C. and Lohr, S. L. (2003), "Robust Estimation of Multivariate Covariance Components," manuscript submitted for publication.

Fay, R.E. (1987), "Application of Multivariate Regression to Small Domain Estimation," in R. Platek et al. (eds.), *Small Area Statistics,* New York: Wiley, 91-102.

Fay, R. E. and R. A. Herriot (1979), "Estimates of Income for Small Places: An Empirical Bayes Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, 78, 269-277.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Fuller, W. A. (1990), "Prediction of True Values for the Measurement Error Model," in P. J. Brown and W. A. Fuller (eds.), *Statistical Analysis of Measurement Error Models and Applications, Contemporary Mathematics Vol. 112,* Providence, RI: AMS, 41-57.

Gleser, L. J. (1981), "Estimation in a Multivariate 'Errors-in-Variables' Regression Model: Large Sample Results," *Annals of Statistics,* 9, 24–44.

Lindley, D.V. (1947), "Regression Lines and the Linear Functional Relationship," *Journal of the Royal Statistical Society Supp.,* 9, 218–244.

Lohr, S. (2003), "Privacy and Survey Research: Ethical and Legal Questions from a Researcher's Perspective," *Bulletin of the International Statistical Institute, Proceedings of the 54th Session.*

Lohr, S. and N. G. N. Prasad (2003), "Small Area Estimation with Auxiliary Survey Data," to appear in *Canadian Journal of Statistics.*

Lohr, S. and J. N. K. Rao (2000), "Inference in Dual Frame Surveys," *Journal of the American Statistical Association,* 95, 271–280.

Lohr, S. and J. N. K. Rao (2002), "Estimation in Multiple Frame Surveys," to appear in *Proceedings of the International Conference on Recent Advances in Survey Sampling.*

Lohr, S. and L. M. R. Ybarra (2003), "Area-level Models using Data from Multiple Surveys," to appear in *Proceedings of Statistics Canada Symposium 2002.*

Madans, J. H., T. M. Ezzati-Rice, M. Cynamon and S. J. Blumberg (2001), "Targeting Approaches to State-Level Estimates," in M.L. Cynamon and R.A. Kulka (eds.) *Proceedings of the Seventh Conference on Health Survey Research Methods*, Hyattsville, MD, Department of Health and Human Services, 239–245.

Maltz, M. (1999), *Bridging Gaps in Police Crime Data*, NCJ Report 176365, Washington, D.C.: Bureau of Justice Statistics.

National Research Council (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond.* C. F. Citro and R. T. Michael, eds. Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics. Washington, D.C.: National Academy Press.

Rao, J. N. K. (2003), *Small Area Estimation,* New York: Wiley.

Schalk, K. L. (2003), *Imputation of Missing Uniform Crime Report Data,* unpublished M.S. thesis, Arizona State University.

Skinner, C. J. and J.N.K. Rao (1996). "Estimation in Dual Frame Surveys With Complex Designs," *Journal of the American Statistical Association*, 91, 349–356.

Wiersema, B., McDowall, D. and Loftin, C. (2000), “Comparing Metropolitan Area Estimates of Crime from the National Crime Victimization Survey and Uniform Crime Reports,” Paper presented at the Joint Statistical Meetings, Indianapolis, August 2000.

Ybarra, L. M. R. (2003), *Area-level Models Using Data from Multiple Surveys*, unpublished Ph.D. dissertation, Arizona State University.

Ybarra, L. M. R. and Lohr, S. (2002), “Estimates of Repeat Victimization Using the National Crime Victimization Survey,” *Journal of Quantitative Criminology,* 18, 1–21.

Session 6

Benefits and Challenges of the Funding Opportunity

General Discussion: Benefits and Challenges of the Funding Opportunity

Brian A. Harris-Kojetin
U.S. Office of Management and Budget

I find myself in the unenviable position of attempting to add something to what has already been expressed here today and trying to follow Fritz, who is always a tough act to follow. Fortunately, my task is not to contribute substantively to the technical discussion, but rather to discuss the benefits and challenges of the Funding Opportunity in Survey Research. Certainly some of the benefits of this program have become quite obvious to everyone here today.

I will briefly address what I see as some of the major benefits of the Funding Opportunity, and then talk about some challenges and opportunities we have in the Federal statistical system that future proposals may help us address.

Benefits

One of the most immediate benefits is the dialogue that we have had here today and the interactions we will continue to have. Certainly one of the goals of this enterprise is to foster greater interaction of the Federal statistical and academic communities about topics of mutual interest and concern. We Feds can benefit and learn about innovations in other sectors, while they learn more about the applied problems we are faced with. Our hope is to push their interest and thinking into areas and applications related to work they are already pursuing, or perhaps even to spark an interest in a new area of research that would be of real benefit to us.

In the process we hope to foster more long-term benefits and lasting relationships that may lead not only to the potential for agencies funding further work but also to new ideas for additional projects and fostering student knowledge of issues and opportunities in the Federal statistical system. There are many interesting problems and challenges we face in Federal statistics, and we need to attract the future talent to deal with these. In some sense, this program becomes one way for us to advertise ourselves and our issues to the next generation through their faculty mentors. NSF considers student support in its decisions and many Federal agencies strongly support this as well.

Looking at the structure of the program itself, the Funding Opportunity provides a valuable mechanism for multiple agencies that don’t have access to a grants process and may not have broad contact with academics and others working on similar issues. Although statistical agencies frequently contract out data collection and perhaps some related methodological and statistical research, most agencies do not on their own have the ability to fund investigator-initiated grants, and could not, on their own muster the resources necessary to fund, manage, and maintain such a program.

The Funding Opportunity is an excellent example of cooperation among statistical agencies for the greater good of the whole Federal Statistical System. Since coming to OMB, I have seen that coordinating and effectively communicating across more than 70 agencies that do some kinds of statistical work is an enormous task. Statistical agencies have frequently been at the forefront of innovation and coordination across government, and the Funding Opportunity is an excellent example of effective use of government resources.

Much of the credit for this success goes to the efforts of Monroe Sirken and Nancy Kirkendall and the collaboration they established between the ICSP and the Federal Committee on Statistical Methodology (FCSM). The FCSM, which consists of about two dozen senior statisticians, methodologists, economists, and managers, is an excellent platform for promoting this research program. The FCSM is perhaps best known for our Working Paper series, which are available on our web site: http:\\www.fcsm.gov and biennial conferences on statistical policy (last November) and the upcoming Research Conference this November. In addition to these forums, the committee is currently engaging in new efforts to reach out and facilitate communication and sharing of expertise across agencies, and to leverage the experience across agencies to provide technical assistance to the Federal statistical system as a whole, and smaller agencies in particular. We are also striving to involve more agencies in the Funding Opportunity to be able to expand the scope and number of projects that we can fund as well as improve communication and collaboration within the statistical system.

This morning Monroe described the history of the Funding Opportunity, and I would like to take a few minutes to talk about its future. The Interagency Council on Statistical Policy (ICSP) has embraced the recommendations of the FCSM research subcommittee to continue funding this program for the next three years. The funding formula was altered slightly to a tiered structure to enable more agencies to contribute in line with their means, but we will achieve approximately the same overall total as before, with some agencies contributing a little more and some a little less. I want to point out that this sets a minimum base of funding. In the past, some agencies have contributed additional money to specific projects that were of direct interest and benefit to them and several have expressed similar sentiments this year.

Unfortunately, for 2003, our description of the Funding Opportunity was not included in time to make NSF's main announcement. Consequently, this fiscal year we saw fewer relevant projects, but still enough of high quality and interest that we anticipate funding. I think it is a real credit to the agency heads and an indication of their commitment that they voted to contribute to this year's program even though we did not have the announcement and did not receive as many proposals as previous cycles. It would have easy for them to opt out this year and perhaps harder to get started back again next year. We strongly encourage you to reach out to colleagues to apply next year and future funding cycles to keep this program alive and vital. The long-term future of this program will be driven by the quality of the projects we are able to fund and the contributions they make to the Federal statistical system.

Challenges and Opportunities

Although I'm very optimistic about the future of this program, I think it's important to balance this with some appropriate cautions. It's not likely that agencies will be feeling that they have much extra room in their budgets in the next few years, so the kinds of projects that are funded and their results will likely impact the long-term future of this endeavor. One strength of the program is that it draws funds from many agencies; however, there are also potential drawbacks when there are diverse stakeholders with different interests and needs. If we fund only projects that appeal to the majority of agencies, there may be some that will be consistently left out, and over time will not see the payoff for their continued participation in the program. We need to carefully consider the breadth of the proposals we fund, but clearly we need proposals that cover the diversity of issues facing the many agencies in the Federal statistical system.

To do this, it's not sufficient just to post announcement on the NSF website—although we have clearly seen this is helpful through an unintended natural experiment. But we also need to reach out to colleagues at forums such as ASA, AAPOR, ISI, AAAS, and other venues to encourage them to apply their expertise to Federal problems or try to help us with issues we are facing. I think we have the opportunity to make this program thrive and involve a growing number of Federal statistical agencies. The challenge will be to attract a diversity of quality projects that will meet a wide variety of needs across these agencies.

Promising Areas

I can't resist this opportunity to put in a pitch for proposals to deal with what I see as some of the pressing problems we face in the Federal statistical system. This is certainly not a representative sample, nor are any of these really surprising.

You didn't need to attend the AAPOR meeting a few weeks ago, to know that there are real concerns about the future of telephone survey methodology, and RDD surveys in particular. However, if you attended the conference, you certainly couldn't have missed the focus on response rates, much of it focused on RDD surveys. Although RDD surveys are not the mainstay of Federal government survey data collections in most statistical agencies, there are a number of Federal RDD surveys across a variety of departments that provide critically important information that is tied to policy making and, in some cases, even quite directly to the distribution of government funds. In addition to concerns about response rates, issues of coverage and the growing impact of cell phones constitute an evolving landscape we need to understand and deal with. We have funded one project in this area, you will hear about at the next seminar, but more work is needed in this area.

There are also growing uses of electronic data collection, spurred on by the Government Paperwork Elimination Act (GPEA). More and more frequently this means using the internet for data collection, rather than just CATI or CAPI. There are certainly many promises in using the web for data collection, but there are also perils. Some government web surveys have certainly been featured in presentations and classes that Roger and his colleagues use to illustrate "what not to do." In addition to the nuts and bolts of doing better web surveys, there are certainly some situations where these kinds of applications are appropriate and others where they are not, at least not as the only mode. A broader framework backed by empirical results to enhance understanding and guide decision-making on how and when to use web-based collections and how to deal with issues of response rates, coverage, and respondent preferences would be very useful to agencies. As Cleo noted in her remarks, web and paper do often need to work together.

Another area I think many statistical agencies would like to see more attention given to is establishment surveys, which I will define quite broadly to include not only private businesses, but also institutions such as hospitals and schools. Compared to demographic or household-based surveys, establishment surveys tend to receive less attention outside government, yet there are many of the general issues are the same with specific variations for this context. For example, we see increasing efforts being required to do effective recruiting and handle nonresponse. Part of the problem is that it is often requiring more and more levels of approval to obtain cooperation. In local areas, we are seeing school districts requiring approval before allowing us to contact schools. Presentations to school boards and even local IRBs are becoming more common. Increased testing required by the No Child Left Behind legislation is, of course, putting more burdens on schools, and our research studies may suffer for it. Likewise, businesses have long complained of the burdens of multiple collections from different agencies. It strikes me, perhaps naively, that a principal investigator funded under this program could learn a great deal about the burdens being placed on organizations, and could perhaps provide some valuable insights that will aid us when agencies are allowed to share data, such as sample frames, and even in cases where they are not.

Finally, confidentiality concerns don't appear to be diminishing anytime soon. In addition to the good technical work on disclosure limitation that has been funded, we could also use more work on respondent perceptions of confidentiality: both from the household and establishment perspectives. In particular, it would be helpful to understand respondent's perceptions of use of data by outside researchers. How much do we tell respondent's about this possibility and how do we tell them to make it clear the protections we have built in? We now have new legislation, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) that allows us to provide a consistently high level of protection to statistical data gathered under a pledge of confidentiality. We need to be able to effectively communicate this to our respondents and make them feel more assured rather than more concerned as some previous research has shown.

To conclude, I think the need for the Funding Opportunity is greater than ever, and I appreciate your active participation. I would like to thank not only the authors and discussants for their excellent work and thoughtful remarks, but also those sponsors who made this possible, and all of you for attending and carrying this back to your agencies, and out to your colleagues. I look forward to the next round of proposals and our next seminar.

Reports Available in the Federal Committee on Statistical Methodology's Statistical Policy Working Paper Series

1. ***Report on Statistics for Allocation of Funds***, 1978 (NTIS PB86-211521/AS)
2. ***Report on Statistical Disclosure and Disclosure-Avoidance Techniques***, 1978 (NTIS PB86-211539/AS)
3. ***An Error Profile: Employment as Measured by the Current Population Survey***, 1978 (NTIS PB86-214269/AS)
4. ***Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics***, 1978 (NTIS PB86-211547/AS)
5. ***Report on Exact and Statistical Matching Techniques***, 1980 (NTIS PB86-215829/AS)
6. ***Report on Statistical Uses of Administrative Records***, 1980 (NTIS PB86-214285/AS)
7. ***An Interagency Review of Time-Series Revision Policies***, 1982 (NTIS PB86-232451/AS)
8. ***Statistical Interagency Agreements***, 1982 (NTIS PB86-230570/AS)
9. ***Contracting for Surveys***, 1983 (NTIS PB83-233148)
10. ***Approaches to Developing Questionnaires***, 1983 (NTIS PB84-105055)
11. ***A Review of Industry Coding Systems***, 1984 (NTIS PB84-135276)
12. ***The Role of Telephone Data Collection in Federal Statistics***, 1984 (NTIS PB85-105971)
13. ***Federal Longitudinal Surveys***, 1986 (NTIS PB86-139730)
14. ***Workshop on Statistical Uses of Microcomputers in Federal Agencies***, 1987 (NTIS PB87-166393)
15. ***Quality in Establishment Surveys***, 1988 (NTIS PB88-232921)
16. ***A Comparative Study of Reporting Units in Selected Employer Data Systems***, 1990 (NTIS PB90-205238)
17. ***Survey Coverage***, 1990 (NTIS PB90-205246)
18. ***Data Editing in Federal Statistical Agencies***, 1990 (NTIS PB90-205253)
19. ***Computer Assisted Survey Information Collection***, 1990 (NTIS PB90-205261)
20. ***Seminar on Quality of Federal Data***, 1991 (NTIS PB91-142414)
21. ***Indirect Estimators in Federal Programs***, 1993 (NTIS PB93-209294)
22. ***Report on Statistical Disclosure Limitation Methodology***, 1994 (NTIS PB94-165305)
23. ***Seminar on New Directions in Statistical Methodology***, 1995 (NTIS PB95-182978)
24. ***Electronic Dissemination of Statistical Data***, 1995 (NTIS PB96-121629)
25. ***Data Editing Workshop and Exposition***, 1996 (NTIS PB97-104624)
26. ***Seminar on Statistical Methodology in the Public Service***, 1997 (NTIS PB97-162580)
27. ***Training for the Future: Addressing Tomorrow's Survey Tasks***, 1998 (NTIS PB99-102576)
28. ***Seminar on Interagency Coordination and Cooperation***, 1999 (NTIS PB99-132029)
29. ***Federal Committee on Statistical Methodology Research Conference (Conference Papers)***, 1999 (NTIS PB99-166795)
30. ***1999 Federal Committee on Statistical Methodology Research Conference: Complete Proceedings***, 2000 (NTIS PB2000-105886)
31. ***Measuring and Reporting Sources of Error in Surveys***, 2001 (NTIS PB2001-104329)
32. ***Seminar on Integrating Federal Statistical Information and Processes***, 2001 (NTIS PB2001-104626)
33. ***Seminar on the Funding Opportunity in Survey Research***, 2001 (NTIS PB2001-108851)
34. ***Federal Committee on Statistical Methodology Research Conference (Conference Papers)***, 2001 (NTIS PB2002-100103)
35. ***Seminar on Challenges to the Federal Statistical System in Fostering Access to Statistics.*** 2004 Forthcoming.

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161; telephone: 1-800-553-6847. The Statistical Policy Working Paper series is also available electronically from FCSM's web site <*http://www.fcsm.gov*>.